Genomic, patterns of selection and differentiation in African populations and implications for mapping disease association

Thesis

UNRESTRICTED

Genomic patterns of selection and differentiation in African populations and implications for mapping disease association.

Abier Elzein

A Dissertation submitted for the degree of Doctor of Philosophy

The Open University

Sponsoring Establishment

Weatherall Institute of Molecular Medicine
The Wellcome Trust Centre for Human Genetics (WTCHG), Oxford.

May 2009

Submission date: 26 May 2009
Date of award: 20 October 2009

PAGE

MISSING

IN

ORIGINAL

Acknowledgments

This work was carried out at the Wellcome Trust Centre for Human Genetics, Oxford between the years 2004 and 2009.

I am greatly indebted to Professor Dominic Kwiatkowski for allowing me to work in his laboratory. Working with him has been both a privilege and an inspiration. I truly have the highest respect for him and for what he chose to make his life about, for his belief that through his enterprise and endeavour the grander good can be achieved. He has been nothing but courteous, kind, and encouraging to me throughout the period of my study and I genuinely thank him for that.

I am immensely grateful to Professor Muntaser Ibrahim for his belief in me. I'm indebted to him for presenting me with my first opportunity to go into research. I have great appreciation for his originality, and excitement about new ideas. I extend my thanks to the members of the IEND and the participants from the villages of Koka and Salala who made this work possible.

A great deal of thanks is also owed to Dr Kirk Rockett for his supervision, help and guidance. I thank him for his graciousness, patience, and generosity with his time. I extend my gratitude to the other members of the Kwiatkowski group, past and present, for their help over the course of this thesis. Thanks especially to Anna Richardson for her help with the laboratory work. Thanks to all the members of the core facilities for running the plates for the SEQUENOM.

A special thanks and acknowledgement goes to those involved in the MalariGen project. I was privileged to have been given early access to and allowed to work with the Gambian dataset in order to test the applicability of my analysis. I am especially thankful to the members of the analysis group Kerrin Small, YY Teo, and Taane Clark for their help and advice.

Among the many who have shared with me their insights and experiences, I am particularly grateful to Julian Forton, Andrew Fry for the numerous stimulating discussions. A special thank you to Bert Mohr for his readily offered and demonstrated friendship, his generous conversations and helpful comments. Very special thanks to Susana Campino, Anita Ghansah, and Elham SadighiAkha for their companionship, friendship, moral support, and all the laughter!

Finally, I would like to express my love and gratitude to my family for their unconditional love and for allowing me to take them for granted.

Genomic patterns of selection and differentiation in African populations and implications for mapping disease association.

Thesis Abstract:

The main objective of this thesis is to gain a better understanding of genomic patterns of natural selection and population differentiation in Africa, where there is great genetic diversity, and of the implications for genetic mapping of complex diseases.

I began by studying two neighbouring villages in eastern Sudan that are of different ethnicity, Hausa and Masalit, and that appear to have different susceptibility to malaria and visceral leishmaniasis (VL). Specifically, I investigated patterns of linkage disequilibrium (LD) and haplotypic signals of positive selection in the 5q31 genomic region which contains immune genes that have been implicated in susceptibility to malaria and VL.

In my first analysis, by genotyping 34 single nucleotide polymorphisms (SNPs) in the 5q31 region, I did not find signals of selection or population differentiation between the Hausa and Masalit using available statistical methods. I conceived the idea that patterns of LD might provide a more sensitive test of population differentiation, and I developed an approach for this using permutation analysis. This method revealed differentiation between the Hausa, the Masalit and other African ethnic groups.

To better understand signals of selection, I next studied a region of the genome associated with a known malaria resistance factor, the haemoglobin S (HbS) variant of the *HBB* gene. By genotyping 26 SNPs in the region of the *HBB* gene, I observed a haplotype that extended in excess of 1 Mb, despite being at high frequency and spanning several recombinational hotspots. This long haplotype carried the HbS allele but, importantly, it could be readily detected without typing the HbS variant.

Building on this observation, I designed a new method to screen the whole genome for long haplotypes that might be signals of selection, and developed a software programme to implement this method. I validated this method using haplotypic data for the Yoruba generated by the HapMap project and complemented by additional SNP data that I generated on HapMap cell lines, and found that the HbS allele resides on a haplotype that extends to 1.2 Mb, and is at strikingly high frequency compared to other haplotypes of similar length on the same chromosome.

Next I applied this method to a large family-based association study of severe malaria in The Gambia, and identified several novel genomic regions with unusually long haplotypes of high frequency. These included a number of regions that may be associated with resistance to severe malaria, and which merit further investigation.

Overall objectives:

- Explore genetic diversity patterns in the Hausa and Masalit of Eastern Sudan, in order to inform the design of future association studies to be carried out in these populations.

- Introduce new approaches and methods to discern genetic differentiation between populations, and to look for signals of natural positive selection in the genome.

- Characterize a set of regions in the genome where malaria selective pressure might have played a role, and highlight them for further future exploration.

List of Abbreviations

°C	degrees Centigrade
µg	microgram
µL	micro litre
µM	micro molar
3'	3 prime end of the sequence
5'	5 prime end of the sequence
AFLP	Amplified Fragment Length Polymorphism
ARMS	Amplification Refractory Mutation System
ATP	adenosine triphosphate
BF	Blood Film
BMI	Body Mass Index
bp	base pair
BSA	bovine serum albumin
CAG	Community Advisory Group
CAR	Central African Republic
CD36	Cluster of Differentiation 36
CD40	Cluster of Differentiation 40
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	Utah residents with ancestry from northern and western Europe
CHB	Han Chinese population of Beijing
CI	confidence interval
cm	centimetre
cM	centiMorgan
CM	Cerebral Malaria
CNV	copy number variation
CO_2	carbon dioxide
CSF2	Colony-Stimulating Factor 2
CTLA4	Cytotoxic T-Lymphocyte Antigen 4
CTP	cytidine triphosphate
Da	Dalton
ddNTP	dideoxynucleotide triphosphate
dl	decilitre
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DZ	Dizygous twins
EBV	Epstein-Barr Virus
EDTA	ethylenediaminetetraacetic acid
EHH	Extended Haplotype Homozygosity
EM	expectation maximization
EMM	entropy maximization method
FBAT	Family Based Association Testing
FDR	False Discovery Rate
Fst	Wright's Fixation Index
g	gram
G6PD	Glucose-6-phosphate Dehydrogenase Deficiency

gDNA	genomic DNA
GMCSF	Granulocyte-macrophage colony-stimulating factor
Gst	The Nei coefficient of differentiation
GTP	guanosine triphosphate
GWA	genome-wide association
HAZ	Z-scores for height-for-age
HbAS	heterozygous form of sickle cell gene
HBB	β-globin gene
HbC	Haemoglobin C
HbE	Haemoglobin E
HBE	Haemoglobin E gene
HbF	Haemoglobin F (foetal haemoglobin)
HbS	Haemoglobin S
HbSS	homozygous form of sickle cell gene
HFE	Hereditary hemochromatosis gene
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
hME	homogenous MassExtend
Hs	Nei within-population diversity measure
htSNPs	haplotype tagging SNPs
HWE	Hardy-Weinberg equilibrium
IFNγ	interferon gamma
IgE	Immunoglobulin E
IgG	Immunoglobulin G
iHS	Integrated Haplotype Score
IL12	Interleukin-12
IL13	Interleukin-13
IL3	Interleukin-3
IL4	Interleukin-4
IL5	Interleukin-5
IL9	Interleukin-9
IRF1	Interferon Regulatory Factor 1
JPT	People of Japanese ancestry from the Tokyo area
K	Kosman expected within-population diversity measure
kb	Kilo base
KB	The Kosman Distance (between-populations diversity measure)
KEGG	Kyoto Encyclopaedia of Genes and Genomes
km	kilometre
KW	Kosman Diversity within-population diversity measure
LCT	Lactase gene
LD	linkage disequilibrium
LDSS	LD-summary statistic
LRH	Long Range Haplotype
M	Molar
MAF	Minor Allele Frequency
MalariaGEN	Malaria Genomic Epidemiological Network
MALDI-TOF MS Spectrometry	Matrix-Assisted Laser-Desorption/Ionosation Time-of-Flight Mass
Mb	Mega base
MBL	Mannan-binding lectin

MCMC	Markov-chain Monte Carlo
MDA	Multiple Displacement Amplification
mg	milligram
$MgCl_2$	magnesium chloride
MHC	Major Histocompatibility Complex
min	minutes
mL	millilitre
mM	milliMole
MRC	Medical Research Council
mtDNA	mitochondrial DNA
Mu	Muller within-population diversity measure
MZ	Monozygous twins
N	The Nei distance (between-populations diversity measure)
NCBI	National Center for Biotechnology Information
ng	nano gram
NH_4	ammonium
NRAMP	Natural Resistance-Associated Macrophage Protein
PAF	Population-Attributable Fraction
PCR	Polymerase Chain Reaction
PCV	Packed Cell Volume
PEP	Primer Extension Preamplification
PGS	Pseudo-Gibbs Sampler
PKDL	Post- Kala-azar Dermal Leishmaniasis
QC	Quality Control
R	The Rogers distance (between-populations diversity measure)
REHH	Relative Extended Haplotype Homozygosity
RFLP	Restriction Fragment Length Polymorphism
rpm	rounds per minute
RVTH	The Royal Victoria Teaching Hospital
SAP	Shrimp Alkaline Phosphatase
SCA2	Spinocerebellar ataxia type 2 gene
sec	second
Sh	Shannon normalized within-population diversity measure
Si	Simpson within-population diversity measure
SMA	Severe Malarial Anaemia
SNP	Single Nucleotide Polymorphism
tagSNP	tagging SNP
TE	Tris-EDTA buffer (10mM Tris-HCl pH 7.5, 1mM EDTA)
TGF-β	Transforming growth factor beta
TNFSF5	Tumor necrosis factor ligand superfamily member 5
TNF-α	Tumour Necrosis Factor alpha
TTP	thymidine triphosphate
U	Unit
UPGMA	Unweighted Pair-Group Method with Arithmetic mean)
UV	Ultra Violet
VL	Visceral Leishmaniasis
WAZ	Z-scores for weight-for-age
WHO	World Health Organization
WHZ	Z-scores for weight-for-height
WTCCC	Wellcome Trust Case Control Consortium

WTCHG	Wellcome Trust Centre for Human Genetics
YRI	Yoruba people of the Ibidan Peninsula in Nigeria

Table of Contents

Chapter 2 – Study Samples, Materials, and Methods.

Chapter 3 - Genetic diversity, LD, and Haplotype structure of the 5q31 region in two ethnically distinct populations from neighboring villages in Eastern Sudan.

List of Figures

Chapter 6:

Chapter 7:

List of Tables

Chapter 1:

Introduction

1.1. Demography of human populations

The evolution of modern human populations has been accompanied by dramatic changes in environment and lifestyle. In the last 100,000 years, behaviourally modern humans have spread from Africa to colonize most of the globe. Humans appear to have experienced a rather strong reduction in the effective population size at the time of migration out of Africa (Garrigan and Hammer 2006).

In that time, humans have been forced to adapt to a wide range of new habitats and climates. Following the end of the last ice age, 14,000 years ago, there was a major warming event that raised global temperatures to roughly their current levels. Further dramatic changes occurred with the transition from hunter-gatherer to agricultural societies, starting about 10,000-12,000 years ago in the Fertile Crescent, and a little later elsewhere. This was also a period marked by rapid increases in human population densities. Increased population density promoted the spread of infectious diseases, as did the new proximity of farmers to animal pathogens (Diamond 2002).

Real populations are rarely simple, so it is difficult to research and develop theories about them. Natural populations are also dynamic in many dimensions: over time they change in size, density and location, and over space they can fragment into several populations and join with others. Our species has had a complex demographic history and provides examples of many kinds of population structure. Also, a considerable part of modern medical genetics

relies on an understanding of human demographic history, so there is a strong demand for high-quality human population-genetic research.

1.2. The HapMap project and other publicly available genome-wide data

Invention of efficient methods for sequencing DNA in the mid 1970s by Maxam and Gilbert (Maxam and Gilbert 1977) and Sanger (Sanger, Nicklen et al. 1977) followed by automation and subsequent development of more and more rapid applications of these methods soon led to the sequencing of complete genomes of a variety of species, and culminated in the completion of the first draft of the human genome in 2001 (Lander, Linton et al. 2001; Venter, Adams et al. 2001). Results of the studies like these have meant a real revolution in the whole of biology, and consequently the present time in biology is commonly called the post-genomic era. Biology and its applied sciences, including medicine, have experienced a radical shift in the methods used, and most of the central problems of biology, like evolution in particular, can now be approached from a completely new quantitative perspective.

Analysis of the human genome sequence has assisted analysis of human diversity, revealing large numbers of single-nucleotide polymorphisms (SNPs), and a structured pattern in the distribution of polymorphisms along chromosomes, whereby common combinations of SNPs, or haplotypes, are observed within populations (reviewed in (Ardlie, Kruglyak et al. 2002)).

The International HapMap Project was founded in 2002, with the goal of mapping the structure of allelic association across the human genome. With the participation of funding agencies, academic research centres, and industrial partners in many countries, the initial aim was to genotype one SNP every 5 kb in the human genome across 270 individuals from four geographical populations. About 1 million SNPs were typed by the completion of phase 1 of

the project and a total of about 4 million SNPs have been typed across these genomes so far after phase 2, providing an unprecedented view of human genetic diversity. The data, with associated summaries and query-based tools, are available online at http://www.hapmap.org .

The aim of the project was to provide a resource that facilitates the design of efficient genome-wide association studies, through characterising patterns of genetic variation and linkage disequilibrium (LD), and facilitating the economic selection of marker SNPs. A lot of insight into the human genetic variation is gained from data generated by the project (McVean, Spencer et al. 2005).

1.3. Organisation of the human genome

1.3.1. Human genetic variation

After the sequencing of the human genome, focus on human genetic variation has come to the forefront, mainly in the form of single nucleotide polymorphisms (SNPs). The number of single nucleotide polymorphisms (SNPs) is the most precise and best, though not the sole, measure of the general amount of genetic polymorphism in humans.

SNPs are the commonest form of variation in the genome. Comparison of any two chromosomes will generally reveal SNPs at 1.2 kb average intervals across the genome (Altshuler, Pollara et al. 2000). The International HapMap consortium (2005) estimated that every 279th nucleotide pair is polymorphic in the human population, while Crawford et al. (2005) reached an even higher estimate of polymorphism. They estimated that every 180th nucleotide pair would be polymorphic. Therefore, with regards to SNPs, we differ from each other only by 3.6-5.6%. The respective figure for big mammals in general is approximately 10% (Crawford, Akey et al. 2005).

SNP markers have revolutionised the way human genetics and disease are studied. Presently they are a key research material and valuable resource for genetic association with heritable traits. They can also tell us many things about the functional parameters and critical regions of a gene, protein, regulatory element or genomic region.

SNPs are thought to arise as a result of mutational defects during DNA replication. Mutation rate has been estimated to be in the order of a probability of $2x10^{-8}$ per nucleotide per generation (Zuckerkandl 2002).The fate of these mutations is often determined by chance. Some mutations will subsequently be lost in the population, that is, the mutation arises in one individual but is not transmitted to the next generation. Many will randomly drift until they reach a state of equilibrium in the population (encapsulated in the neutral theory (Kimura 1968)), while still others respond to selective pressures, for example if the SNP confers some survival advantage.

The methods by which SNPs arise and are maintained ensure that their demography in human populations is diverse and that their distribution and frequency vary widely with ethnicity and gene region (Salisbury, Pungliya et al. 2003). Most are found in non-coding regions of the genome where they often have little function although a few may regulate gene function. This latter group includes SNPs in the promoters of genes (Hoogendoorn, Coleman et al. 2003), those affecting splice sites (Liu, Cartegni et al. 2001), or SNPs within gene enhancers (Nobrega, Ovcharenko et al. 2003). Within coding regions SNPs can lead to an amino acid substitution potentially affecting protein function, although the majority produce changes that have little or no effect (Cargill, Altshuler et al. 1999).

Presence and frequency of SNPs varies with populations. Younger populations such as Europeans who migrated from Africa some time in the past, taking a subset of the diversity with them (Goldstein and Chikhi 2002), are expected to have fewer and younger SNPs.

These differences in allele frequencies between populations are potential confounders in association studies (Cardon and Bell 2001; Emahazion, Feuk et al. 2001), although they can also be used to map complex traits (admixture mapping) (Collins-Schramm, Phillips et al. 2002).

1.3.2. Linkage disequilibrium

Recently, there has been tremendous interest in empirically establishing the patterns of allelic association, also known as LD, among polymorphic variants of the human genome. When two alleles at adjacent loci co-occur in a chromosomal segment more often than expected if they were segregating independently in the population, the loci are in linkage disequilibrium. The profile of LD depends on the age of the mutations. It is eroded by recombination (Jeffreys, Kauppi et al. 2001), and mutation. Both of these forces act slowly, each occurring at an average rate of about 10^{-8} probability of occurrence per base pair (bp) per generation. Since recombination occurs at successive generations, it will have a greater effect upon the LD between older SNPs than between younger ones. This explains the reduced LD seen in older populations such as Africans (Hull, Ackerman et al. 2001; Reich, Cargill et al. 2001).

Genetic drift and the demographic history of a given population are also some of the major factors that determine LD. The ancestry of a population can affect the extent and level of LD (Goldstein and Chikhi 2002; Rosenberg, Pritchard et al. 2002). Population expansion tends to reduce overall LD, whilst population admixture, inbreeding, migration, and geographical subdivision (bottlenecks), tend to increase LD. However, the inherent complexity of human history means that all of the above can have unique and complicated effects on LD.

Natural selection has unique effects on the allelic architecture of a locus. Directional or positive selection tends to reduce surrounding variation and increase the relatedness of surrounding markers (and consequently LD) as the selected allele increases in the population. Balancing selection tends to produce a number of intermediate frequency alleles and the overall effect on LD can be quite complex, depending upon the number of balancing alleles and the strength of selection at each one.

Gene conversion (Ardlie, Liu-Cordero et al. 2001) also determines the quantity and quality of LD, where a short segment of one chromosome is transferred to another. The frequency and determinants of this phenomenon are not well known, although it has been speculated that widespread gene conversion events are responsible for the poor LD sometimes observed over short distances, occasionally even with adjacent makers.

Over many generations LD in a region is uniquely shaped by all these forces and in practice, LD is a window to the ancestry of a group of markers.

1.3.3. LD organisation in the human genome

With the sequencing of the human genome and development of high-throughput genomic methods, it became clear that LD is more varied across regions, and more segmentally structured (Daly, Rioux et al. 2001; Gabriel, Schaffner et al. 2002), than had previously been supposed.

Recombination rates typically vary dramatically on a fine scale, with hotspots of recombination explaining much crossing over in each region (Jeffreys, Kauppi et al. 2001). The generality of this model has recently been demonstrated through computational methods that allow estimation of recombination rates (including hotspots and coldspots) from genotype data (Crawford, Bhangale et al. 2004; McVean, Myers et al. 2004). It is estimated

that about 80% of all recombination has taken place in about 15% of the Sequence (HapMap 2005). With most human recombination occurring in recombination hotspots, the breakdown of LD is often discontinuous. A 'block-like' structure of LD is manifest by segments of consistently high LD that break down where high recombination rates or recombination hotspots cluster. Despite the initial promise of haplotype blocks, more in-depth views of haplotype blocks suggest that the original belief that haplotype blocks represent a fundamental aspect of the human genome appears to be an oversimplified view of genome organization (Phillips, Lawrence et al. 2003; Ke, Hunt et al. 2004). While the idea of blocks appears to be true on a large scale, fully defining blocks, and block boundaries in particular, is far less robust, and appears to be at best an approximation of the complete picture.

1.3.4. Usefulness of studying LD

Defining regional LD patterns could play an integral role in deciphering mixed genetic signals. For example, consider two different SNPs in two juxtaposed genes separated by some distance, genotyped in two independent but identical studies, in which both SNPs are associated with a disease. Knowing the inherent LD between the two markers and their relation to other surrounding markers might make it possible to determine whether there exist two independent associations, whether the associations are related, and whereabouts the causative marker could be. Therefore it may be possible to use the pattern of LD to help define the boundaries of disease associations (Tiret, Poirier et al. 2002), provide a greater confidence in results, and reduce genotyping redundancy.

The extent of LD across genomic regions is a crucial parameter for defining the statistical power of association studies utilizing single nucleotide polymorphisms (SNP) as surrogate

genetic markers (Schork 2002), and for guiding the selection and spacing of such polymorphisms to create marker maps useful in candidate gene, candidate region, and whole-genome association studies (De La Vega, Dailey et al. 2002). The feasibility of such an exercise relies upon how much redundancy exists in the genome, that is, if one marker is well associated with another there is little point in testing both. If the genome could easily be broken into haplotype blocks, genome-wide SNP associations would be much more realistic, as a small subset of markers would provide most of the information about relatively long genetic stretches. This idea has prompted some researchers to approach genome scanning using only those SNPs that define haplotype blocks, so called tagging SNPs (Zhang, Calabrese et al. 2002; Ke and Cardon 2003).

Patterns of LD are also informative about population histories and human migrations (Tishkoff, Dietzsch et al. 1996; Reich, Cargill et al. 2001; Plagnol and Wall 2006), recent natural selection (Sabeti, Reich et al. 2002; HapMap 2005; Voight, Kudaravalli et al. 2006), and the distribution and evolution of recombination hotspots (McVean, Myers et al. 2004; Fearnhead and Smith 2005; Ptak, Hinds et al. 2005).

1.3.5. Haplotype structure of human populations

The particular combination of marker alleles along a chromosome is referred to as a haplotype. LD and haplotypes are, of course, related. The number and structure of haplotypes across a region is a partial surrogate for the level of LD.

The initial belief that haplotype block boundaries and haplotypes were largely shared across populations was a foundation for constructing a haplotype map of the human genome. The HapMap data document the generality of a block-like pattern of LD with regions of low and high haplotype diversity but differences among the populations. The International HapMap

Consortium (2005) observed that the different population groups are characterized by different haplotype frequencies and, to some extent, different combinations of SNPs inside the haplotypes.

Studies of many additional populations demonstrate that LD patterns can be highly variable among populations both across and within geographic regions. Recent studies have reported significant variation among populations in block structure and tagSNPs. Sawyer *et al* (Sawyer, Mukherjee et al. 2005) studied three loci in 16 diverse populations with an emphasis on African and European populations. They found significant quantitative and qualitative variation in LD among populations both across and within geographic groups, and no group showed consistency in patterns of LD for all three loci under study. Liu *et al* (Liu, Sawyer et al. 2004) reached a similar conclusion with respect to tagSNPs.

Gu et al (Gu, Pakstis et al. 2007) observed large variation in block partition among 38 world populations from different regions, and their results also showed that significant variation can occur among populations within geographic regions. None of the block-defining algorithms they used produced a consistent pattern even among populations of similar geographic origins. They attributed such differentiation to be due to the individual population demographic history and the combined effects of genetic factors such as drift, mutation and recombination. At average sizes of 50 individuals in each group, sampling error was considered a minor factor (Fallin and Schork 2000).

The large amount of variation in haplotype block structure among global populations reflects the considerable haplotype variation among these populations seen previously at several loci (Kidd, Pakstis et al. 2004).

The evolution of the human genome seems to have resulted in a mosaic of discrete segments, each with its own unique history and relatedness to different contemporary and ancestral individuals (Paabo 2003).

1.4. African genetic diversity

African populations are characterized by greater levels of genetic diversity, extensive population substructure, and less LD among loci compared to non-African populations. Africans also possess a number of genetic adaptations that have evolved in response to diverse climates and diets, as well as exposure to infectious disease.

A majority of studies have shown that African populations harbour more genetic diversity than non-African populations for mitochondrial DNA (mtDNA) sequences, Y chromosome microsatellites, Y chromosome sequences, Y chromosomal single-nucleotide polymorphisms (SNPs), X chromosome sequences, autosomal microsatellites, autosomal sequences as well as autosomal SNPs (Excoffier 2002). This basic result is generally interpreted as evidence for an African origin of our species.

It was also observed that the amount of human SNP in USA was higher among people of African origin than in people of European origin. When comparing any two individuals, every 1110th nucleotide pair in the mean was different in people of African origin, while the respective figure in the people of European origin was every 1435th. Taking into account that the haploid human genome consists of 3200 million nucleotide pairs of DNA, one can calculate that any two individual genomes differ by 2.89 million nucleotide pairs in the former group, and by 2.23 million nucleotide pairs in the latter group (Crawford, Akey et al. 2005).

Hinds et al. (Hinds, Stuve et al. 2005) studied the distribution of over 1.5 million single nucleotide polymorphisms (SNPs) in 71 Americans of European, African, and Asian ancestry. They found that 93.5% of these SNPs could be observed in the people of African ancestry, 81.1% in people of European, and 73.6% in people of Asian ancestry. African-Americans also had overwhelmingly more (218 500) so called private SNPs, which are segregating in only one population, than European- American (44 500) or Asian-American individuals (25 957).

Studies have shown marked differences in the extent of LD observed between African and non-African populations. De la Vega et al (De La Vega, Isaac et al. 2005) observed that, by all measures used, out-of-Africa populations showed over a third more LD than African-Americans.

The amount and nature of haplotype variation also suggest an African origin. African populations have relatively larger population size over a long history, and thus preserve more haplotypes than non-African populations, which are considered to have experienced bottleneck events and have much smaller population sizes historically. In a survey of haplotype structure across 12 Mb of DNA sequence in 927 individuals representing 52 populations from different parts of the world, it was found that the diversity of haplotypes decreases as distance from Africa increases (Conrad, Jakobsson et al. 2006).

1.5. Population differentiation and structure

Ancient demographic and evolutionary events have left imprints that can be observed in present-day gene differences. Therefore, genetic investigations of present-day human populations can illuminate the evolutionary history of our species.

Given enough time following populations' division, isolated groups become genetically more divergent with time, either by acquiring new mutations, or when the original haplotypic backgrounds on which existing variants lie change in abundance by drifting upward or downward in the population. That plus the reshuffling caused by recombination, affect the associations between these variants in ways that are specific to each population group. These changes can either take place over long time periods due to genetic drift or can happen over a relatively short time span when local environmental factors exert selective pressure on a functional variant sweeping nearby neutral polymorphisms up with it.

Differentiating genetically between populations is valuable for admixture and population stratification detection. As well as lending itself to the analysis of case control association studies - by highlighting any hidden population structure in the sample that might generate spurious results if gone undetected-, discerning populations genetic differentiation might be hugely beneficial in the field of pharmaco-genetics, where the genetic structure of a population is used as a predictor of the efficacy of drugs or the likelihood of adverse reactions (Jorde and Wooding 2004).

In a study of genetic variation among world populations; African populations were found to be more diverse than other continental groups and the largest genetic distance was seen between them and non-African populations (Watkins, Rogers et al. 2003). Nei et al. (Nei 1982) studied the genetic relationships of various races in each group of Europeans, Africans, and Asians, and found all European populations to be genetically close to one another, whereas many African tribes show large extents of genetic differentiation.

1.6. The concept of positive selection, selective sweeps and the marks they leave in the genome

Our ancestors were exposed to new environments and diseases. Those who were better adapted to local conditions passed on their genes, including those conferring these benefits, with greater frequency. This process of natural selection left signatures in our genome that can be used to identify genes that might underlie variation in disease resistance or drug metabolism.

The elimination of variation in regions linked to a recently fixed beneficial mutation is known as a "selective sweep" and has recently been the focus of much theoretical and empirical attention (Fay and Wu 2000; Kim and Stephan 2002; Przeworski 2002; Sabeti, Reich et al. 2002; Jensen, Kim et al. 2005).

This reduction in variation at linked sites is a feature of two types of selection. The first – background selection - removes deleterious mutations and eliminates variation at linked sites. The strength of this effect will vary with the recombination rate, the magnitude of selection and the mutation rate. The second - genetic hitchhiking- predicts that if a mutation increases in frequency in a population as a result of positive selection, linked neutral variation will be dragged along with it. As a consequence, variation that is not linked to the adaptive mutation is eliminated, resulting in a selective sweep. Therefore, models predict that genetic hitchhiking will cause a greater overall reduction in genetic diversity, and that the effect will be more pronounced in regions of lower recombination. Both types of selection will result in an overall positive correlation between genetic diversity and recombination rate if the strength and frequency of positive and/or background selection are sufficiently high throughout the genome (Kaplan, Hudson et al. 1989).

In recent years, there has been a dramatic increase in the use of genome-wide scans to identify adaptively evolving loci in the human genome. A large body of evidence is accumulating to suggest the wide spread of positive selection footprints in our genome, with the availability of large scale, genome-wide, population genetic variation data (Andolfatto 2001).

By contrasting patterns of coding sequence polymorphism, identified by direct sequencing of 39 human individuals for 11 624 protein-coding genes to divergence between humans and chimpanzees, Bustamante et al. (Bustamante, Fledel-Alon et al. 2005) found strong evidence that natural selection has shaped the recent molecular evolution of our species. The analysis discovered 304 (9.0%) out of 3377 potentially informative loci showing evidence of rapid amino acid evolution. Furthermore, 813 (13.5%) out of 6033 potentially informative loci showed a paucity of amino acid differences between humans and chimpanzees, indicating weak negative selection and/or balancing selection operating on mutations at these loci. Bustamante et al. (2005) also found that the distribution of negatively and positively selected genes varies greatly among biological processes and molecular functions. Some classes, such as transcription factors, show an excess of rapidly evolving genes, whereas others, such as cytoskeletal proteins, show an excess of genes with extensive polymorphism within humans and yet little amino acid divergence between humans and chimpanzees.

Positive natural selection as such has been observed in the human lineage in various functional classes of genes. More importantly, the most recent studies suggest that this selection is still going on. In addition to brains, positive selection has been found for genes regulating the development of immune response, reproduction, and sensory perception. The results seem plausible in terms of evolutionary predictions (Sabeti, Schaffner et al. 2006).

Why it is important to look for signals of positive selection in the human genome.

The search for signals of positive selection in the human genome remains one of the most important and challenging areas of research in genetics (Akey, Zhang et al. 2002). Inferences regarding the patterns and distribution of selection in genes and genomes as well as insights into how genes predispose individuals to disease are gained from this kind of genetic analysis. It may provide important functional information that might inform the development of improved therapeutic and disease-prevention strategies.

Searching for signatures of positive selection has recently emerged as a strategy for identifying putative causal determinants of disease pathogenesis, without the requirement for *a priori* experimental determination of function. This strategy is based on the rationale that genetic factors that confer protection from disease should be under positive selective pressure and, thus, display signatures of positive selection.

1.7. Infectious diseases and evolution

Genetic epidemiology seeks to determine whether the variation in disease severity can be accounted for by variation in the genome. For a disease to exert selective pressure it would have to have a significant effect on morbidity and mortality before reproductive age and to have been exerting these effects for long periods of time.

Several common genetic disorders have been associated with protection from infectious disease, suggesting that their continued presence in the population has been the result of selection by infectious agents. A study of almost 1,000 adoptees in Denmark found that the host genetic contribution to susceptibility to premature death from infection was higher than for cancer or cardiovascular disease (Sorensen, Nielsen et al. 1988).

1.8. Investigation of complex diseases

The contribution of DNA to disease is perhaps best illustrated by the many genes that have been implicated in inherited disorders. Traditionally, the primary focus of medical genetics has been monogenetic (single gene) disorders such as cystic fibrosis and haemophilia, where mutations in the gene lead directly to manifestation of the disease phenotype. When a disease trait is monogenic, when inheritance patterns are predictable and when there is full penetrance of effect, correlation of the variation at a given gene locus with a given disease phenotype may be relatively straightforward. Uncovering the genes responsible for these kinds of disorders is primarily done by tracing the segregation of markers within affected families using microsatellite markers across the whole genome (linkage studies) and then using positional cloning to further localize the effect. More than a thousand genes for rare, highly heritable Mendelian disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease.

Most Mendelian disorders are rare and although important, they have relatively little impact on public health. The genetic basis for more common diseases affecting humans has been less straightforward to illicit. These conditions are thought to be the expression of the complex interplay between multiple genetic determinants of disease risk of varying influences that act along side several environmental factors to produce the final phenotype.

In the past, studies of common diseases have fallen into two broad categories: family-based linkage studies across the entire genome, and population- based association studies of individual candidate genes. Although there have been notable successes, progress has been slow due to the inherent limitations of the methods; linkage analysis has low power except when a single locus explains a substantial fraction of disease, and association studies of one

or a few candidate genes examine only a small fraction of the 'universe' of sequence variation in each patient.

Nevertheless, with huge recent advances in molecular biology and computational technology, a genome-wide approach became feasible: The International HapMap resource, which documents patterns of genome-wide variation and LD in four population samples, greatly facilitates both the design and analysis of association studies. There was also the availability of dense genotyping chips, containing sets of hundreds of thousands of single nucleotide polymorphisms (SNPs) that provide good coverage of much of the human genome meant that for the first time genome-wide association (GWA) studies for thousands of cases and controls are technically and financially feasible. Moreover, appropriately large and well-characterized clinical samples have been assembled for many common diseases.

Demonstrating the feasibility of GWA studies, a path breaking Wellcome Trust Case Control Consortium (WTCCC) study (Saxena, Voight et al. 2007) was undertaken in the British population. The study examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals across the seven diseases of a $P < 5 \times 10^{-7}$.

1.9. Challenges of interpreting results from genetic association studies

The practical implementation of genetic association studies has proved more difficult than first thought (Emahazion, Feuk et al. 2001; Ioannidis, Ntzani et al. 2001). Finding the variants that directly influence complex diseases has been problematic, and there has been an inability to replicate many associations. This is a multifaceted problem involving issues of sample size, population stratification, and the complexity of LD structure both within and between populations.

A large case-control study performed in The Gambia showed that HLA-B*5301 and HLA-DR-1*1302 were associated with a reduced risk of severe malaria in childhood (Hill, Allsopp et al. 1991). While in a study performed in Kenya, a significant association of HLA-DR-1*0101 with protection against severe forms of malaria was found but no association with HLA-B*5301 and HLA-DR-1*1302 was found (Riley, Olerup et al. 1992). Several hypotheses have been made to explain this heterogeneity in HLA associations in different countries. One possibility is that the different genetic background of the populations under study could affect the pattern of association. It is also possible that the composition of parasite populations and, therefore, the variation in parasite antigens may exert an effect on the selection of the HLA molecules. Finally, another explanation could involve the different epidemiological context with the influence of many co-infections.

Our poor understanding of many fundamental questions about immunity to malaria – such as the molecular mechanisms that are critical for control of parasite density or elimination of parasites from the bloodstream – is a significant impediment to vaccine development. There is no lack of hypotheses and experimental models, many of which are conflicting. The problem is how to determine the importance of different immune mechanisms in natural human infection. A further problem is how to investigate the many novel immune genes that are believed to exist in the human genome but whose function is currently unknown.

1.10. Background on Malaria

1.10.1. Outline the malaria problem

Malaria is an infection of the blood with parasitic protozoa of the genus *Plasmodium*, which are transmitted to humans through the bites of infected female *Anopheles* mosquitoes. *P. falciparum* causes the majority of infections in Africa and is responsible for most cases of

severe disease and mortality. Malaria has an enormous impact on health worldwide with an estimated 35% of the world's population living in areas where there is some risk of *P. falciparum* transmission, creating a current disease burden of 300-500 million cases, and one to three million deaths each year most of them children (Sachs and Malaney 2002; Guerra, Gikandi et al. 2008). Africa gets the biggest toll of the disease mortality especially in remote rural areas with poor access to health services (WHO report. 2003). From all cases of malaria, over 90% of the total death toll is confined to sub-Saharan Africa (Cot and Deloron 2003; Breman, Alilio et al. 2004; Snow, Guerra et al. 2005). There is a striking correlation between malaria and poverty, with lower rate of economic growth in malaria endemic countries (Sachs and Malaney 2002).

Malaria is endemic in more than 90 countries and, together with HIV, tuberculosis and diarrhoeal diseases, constitutes one of the major causes of death by infectious diseases worldwide (Greenwood and Mutabingwa 2002). The malaria situation in sub-Saharan Africa has continued to deteriorate over the past decade mainly due to the widespread emergence of drug-resistant parasites as well as resistance to insecticides in the *Anopheles* vector.

1.10.2. Best approach to tackle the malaria problem

Because of the inability of most individuals in malaria-endemic regions to access or afford optimal treatment, and the ever-evolving drug resistance, the most effective strategy to reduce the number of malaria-related deaths should be through prevention of malaria disease. In 1998, the Roll Back malaria campaign vowed to halve the global burden of malaria by 2010. As we approach the later stage of this campaign, it appears that the established methods, including case management (diagnosis and chemotherapy) and integrated vector control (insecticide-treated bed nets and residual house spraying), are not

sufficient alone to achieve world-wide reductions in the burden of malaria. Many malaria control experts believe that sustainable reductions in malaria control will be impossible in the absence of a safe and effective vaccine against the disease (reviewed in (Tongren, Zavala et al. 2004)).

But, so far, efforts to develop effective anti-malarial vaccines have remained disappointingly unsuccessful, despite immense research efforts worldwide (Richie and Saul 2002). The challenges originate from the parasite complexity, its ability to change through its life cycle both in the human and in the mosquito, and its ability to hide from the immune system.

A better understanding of the natural mechanisms of host defence against the *Plasmodium* parasite may provide new targets for therapeutic intervention in this disease. Such factors may manifest themselves as genetic determinants of susceptibility to infection in endemic areas of disease or during epidemics (Hill 1998).

1.10.3. The malaria parasite

Members of the genus *Plasmodium* belong to the phylum Apicomplexa. The Apicomplexa are a large group of almost exclusively parasitic protozoa. There are about 150 species of *Plasmodium*, but only four species of *Plasmodium* typically infect humans; *P. falciparum, P. vivax, P. ovale* and *P. malariae.* However, a parasite species which primarily infects monkeys, *P.knowlesi*, has recently been reported in humans in Malaysian Borneo (Singh, Kim Sung et al. 2004).

The four *Plasmodium* species, with the exception of *P. malariae* (which may affect the higher primates) are exclusively human parasites. *P. falciparum* has a worldwide distribution and is concentrated in the tropics and subtropics. It is considered the youngest in evolutionarily terms and the least efficient as a parasite because its malignant nature tends to

eliminate its host (Joy, Feng et al. 2003; Su, Mu et al. 2003; Mu, Awadalla et al. 2005). *P. falciparum* is the most virulent species and is responsible for most malaria-related deaths, especially in Africa (Greenwood, Bojang et al. 2005). From a population genetics perspective, such a virulent parasite serves as a strong selective agent for genetic resistance.

Life cycle of the malaria parasite

Malaria parasites have a complex life cycle which is split between a vertebrate host and an arthropod vector. Vertebrate hosts include reptiles, birds, rodents, monkeys and humans. The arthropod vector of human *Plasmodium* species is the female *Anopheles* mosquito. The *Anopheles gambiae* complex comprises the most important vectors of malaria in sub-Saharan Africa (Coluzzi, Sabatini et al. 1979).

Human infection is initiated when sporozoites are injected with the saliva of a feeding female anopheline mosquito (Figure 1.10.3., C). The sporozoites enter the circulatory system and within 30-60 minutes will invade hepatocytes either directly or via the Kupffer cells. After invading the hepatocyte, the parasite undergoes an asexual replication. Schizogony refers to a replicative process in which the parasite undergoes multiple rounds of nuclear division without cytoplasmic division followed by a budding to form merozoite progeny (Figure 1.10.3., A). Merozoites are released into the circulation following rupture of the host hepatocyte and invade erythrocytes.

After entering the erythrocyte the merozoite develops into trophozoites (Figure 1.10.3., B). In the course of their development, they absorb the haemoglobin of the red blood cells leaving as the product of digestion a pigment called hemozoin, a combination of haematin with protein. After a period of growth the trophozoites undergo an asexual dividing process of erythrocytic schizogony. The nucleus divides 3-5 times into a variable number of small nuclei. This is soon followed by the division of cytoplasm, forming a schizont. When the

infected erythrocyte ruptures, merozoites are released and invade new erythrocytes. This erythrocytic cycle of schizogony is repeated over and over again in the course of infection, leading to a progressive increase of parasitemia until the process is slowed down by the immune response of the host or by the action of effective antimalarial drugs.

As an alternative to schizogony some parasites will undergo a sexual cycle and terminally differentiate into micro- (male) or macro- (female) gametocytes (Figure 1.10.3., B). The precise mechanism for this differentiation remains unknown but it is thought that the development of host antibodies may play a large part in the process. It has been observed that the commitment to gamete formation occurs only after the peak of asexual parasitemia is reached or after drug treatment.

When a female mosquito ingests the blood of a human host with malaria parasites in the circulation, the mature sexual cells undergo a series of developments in the stomach of the mosquito. These gametocytes migrate into the mosquito gut, where exflagellation of microgametocytes occurs, and the macrogametocytes are fertilized (Figure 1.10.3., C). The resulting ookinete penetrates the wall of a cell in the midgut, where it develops into an oocyst. Sporogeny within the oocyst produces many mobile sporozoites. The sporozoites migrate (after rupture of the oocyst) from the body cavity of the mosquito to the salivary glands and the mosquito becomes infective to another host. The cycle continues when these sporozoites are injected into a human host when the mosquito feeds.

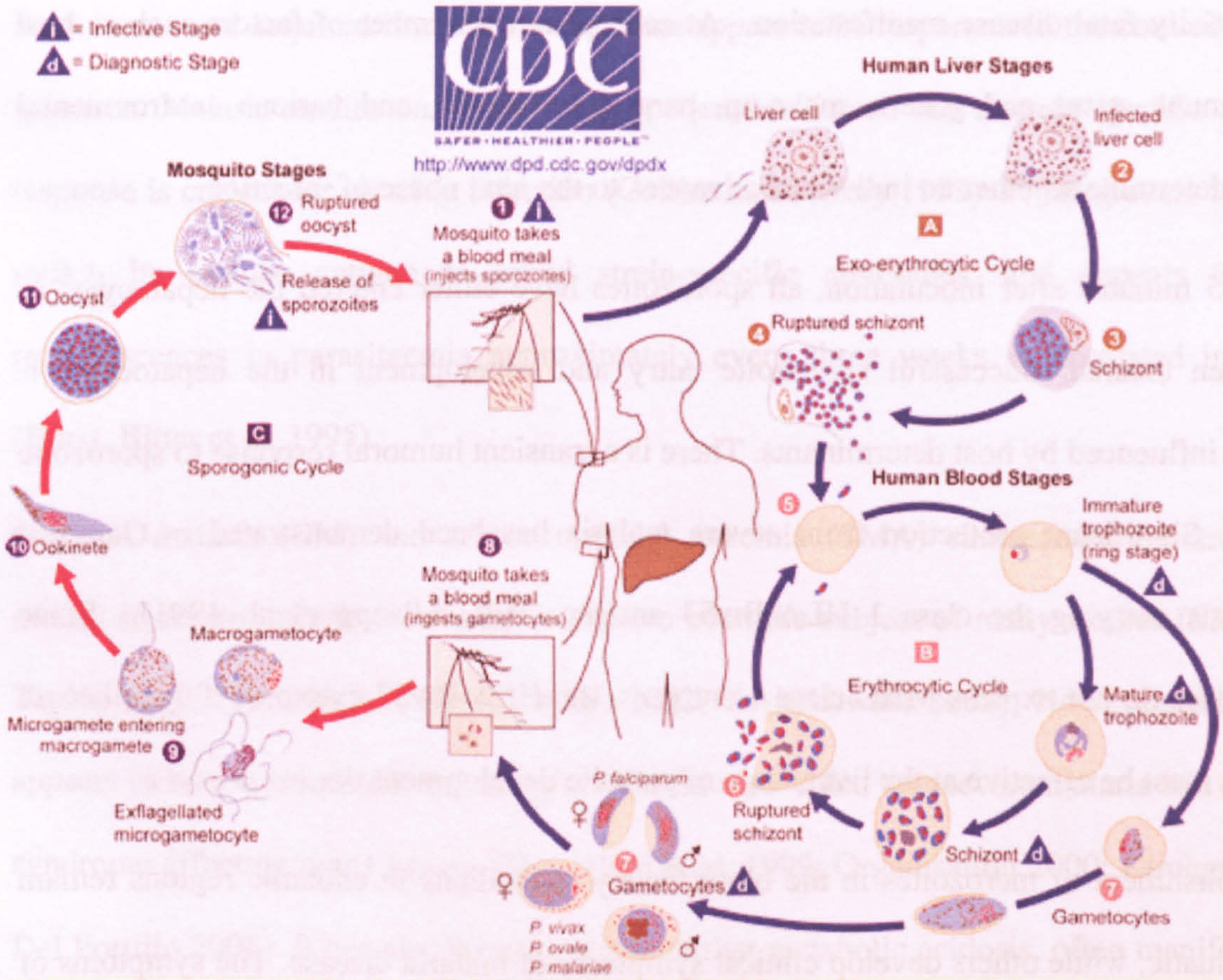


Figure 1.10.3: The life cycle of *Plasmodium falciparum*.
Illustration from CDC's website for laboratory identification of parasites.
(http://www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/Malaria_il.htm).

1.10.4. Clinical Aspects of Malaria

The range of clinical outcomes in response to *P. falciparum* infection is broad, ranging from asymptomatic infection to severe disease and fatality. In malaria-endemic regions of Africa, the majority of individuals carry *P. falciparum* parasites during the high malaria transmission season. While most individuals are asymptomatic or display only mild symptoms of malaria disease, 1-2% of infections progress into severe and potentially fatal complications (Greenwood and Mutabingwa 2002). Approximately, 20% of severely affected individuals die as a result of malaria (White 1987; Pasvol 2006). Several transitional phases, take place in the progression from initial inoculation with *P. falciparum* sporozoites

to potentially fatal disease manifestation. At each phase, a number of factors such as host age, immune status and genetic make-up, parasite virulence, and various environmental factors, determine whether an individual advances to the next phase of infection.

About 45 minutes after inoculation, all sporozoites have either entered the hepatocytes or have been cleared. Successful sporozoite entry and development in the hepatocytes is partially influenced by host determinants. There is a transient humoral response to sporozoite antigens. Significant protection from severe malaria has been demonstrated in Gambian individuals carrying the class 1 HLA-Bw53 antigen (Hill, Allsopp et al. 1991). Since erythrocytes do not express MHC class 1 antigens, the HLA-Bw53 cytotoxic T-lymphocyte response must be effective at the liver stage of parasite development.

On establishment of merozoites in the blood, many individuals in endemic regions remain asymptomatic, while others develop clinical symptoms of malaria disease. The symptoms of clinical malaria include a combination of fever, chills, sweats, headaches, muscular ache, nausea and vomiting, vague abdominal discomfort and lethargy. These symptoms are common to all four species of *Plasmodium* that infect humans. Physical findings in uncomplicated *P.falciparum* malaria may include mild jaundice, enlarged spleen and/or liver, and increased respiratory rate.

The cause of fever is the induction of host cytokines in response to parasite products released into the circulation following schizont rupture. Host cytokine production (and fever) is part of the host immune response to control parasite multiplication, although over-production of some pro-inflammatory cytokines such as tumour necrosis factor-α (TNF-α) may also have adverse effects on the host (Kwiatkowski and Bate 1995). Indeed, a number of polymorphisms in genes encoding cytokines have been associated with severe malaria (reviewed in Kwiatkowski 2005 (Kwiatkowski 2005).

Both innate and acquired immune mechanisms appear to be important in the control of acute infection. Malaria studies on neurosyphilis patients indicated that a strain-specific antibody response is critical for clearing infection (Collins and Jeffery 1999). *P. falciparum* is able to switch its surface antigen to avoid strain-specific antibodies, and appears to cause recrudescences in parasitaemia approximately every three weeks in untreated infections (Borst, Bitter et al. 1995).

Cerebral malaria (CM) and severe malarial anaemia (SMA) characterise the traditional model of severe malaria in children and have been the subject of many studies (Molyneux, Taylor et al. 1989; Miller, Baruch et al. 2002; Kwiatkowski 2005). However, this model appears to be too simplistic and many studies have revealed that severe malaria is a complex syndrome affecting many organs (Day, Hien et al. 1999; Organisation 2000; Kirchgatter and Del Portillo 2005). It has also become apparent that metabolic acidosis, often manifesting as respiratory distress, is an important component of the severe malaria syndrome (Krishna, Waller et al. 1994; Day, Phu et al. 2000). Furthermore, metabolic acidosis has been demonstrated to be the best independent predictor of a fatal outcome in both adults and children (Allen, O'Donnell et al. 1996; Marsh, English et al. 1996). In endemic conditions, functional immunity is acquired early in life and over 75% of mortality affects children less than 5 years of age (Winstanley, Ward et al. 2002). In contrast, areas of infrequent parasites exposure suffer a substantially lower mortality burden. However, functional immunity does not develop as readily and all age groups are equally affected (Smith, Leuenberger et al. 2001; Reyburn, Mbatia et al. 2005).

The clinical outcome of malaria infection depends on several factors with regards to the parasite and host as well as geographic and social factors (Miller, Baruch et al. 2002) (Hill 1999; Kwiatkowski 2005). The combination of these factors determines the outcome of the disease.

1.11. Evidence of presence and nature of genetic susceptibility to malaria

Over the past 50 years, a large body of evidence has accumulated to indicate that genetic variants influence the onset, progression, severity of disease and ultimate outcome of malaria infection in humans. This genetic component is often complex and multigenic (Garcia, Cot et al. 1998), and its analysis by genetic epidemiology, linkage and association studies, as well as by candidate-gene functional studies, has revealed important three-way interactions between host genes, environment and the malaria parasite. In particular, the malarial parasite appears to have exerted positive heterozygote selection for retention of otherwise-deleterious and disease-associated polymorphisms at certain human genes. One of the most complex issues of human genomics in malaria is to quantify the contribution of host genetic determinants in the variation in susceptibility to disease.

Genetic studies of malaria in mice, in which environmental conditions can be stringently controlled, have demonstrated that certain strains of mice consistently show a greater degree of resistance to infection than other strains (Stevenson, Lyanga et al. 1982; Stevenson and Skamene 1985).

In a study to determine the heritability of malaria in Africa, Mackinnon and colleagues (2005) demonstrated that genetic and unidentified household factors each accounted for around one quarter of the total variability in malaria incidence in a Kenyan population. Furthermore, only a small proportion (under 5%) of this variation was attributable to the well known sickle (HbS) and α-thalassaemia malaria resistance genes, indicating the existence of many more host genetic determinants of malaria, each with small effects (Mackinnon, Mwangi et al. 2005).

The first twin study in malaria was performed by Sjoberg et al. who evaluated the genetic contribution in antibody levels against a major malaria antigen (Pf155/RESA) in Liberia and

in Madagascar (Sjoberg, Lepers et al. 1992). Antibody responses to the intact antigen and to some of its immunodominant epitopes were found to be more concordant among monozygotic twin pairs than in dizygotic pairs or siblings living under the same environmental conditions. These results suggested that antibody responses are genetically regulated. Another twin study of malaria infection and disease was performed on 258 Gambian twins comprising 40 monozygous (MZ) twins, 217 dizygous (DZ) twins and one pair of unknown zygosity by Jepson et al (Jepson, Banya et al. 1995). They found that concordant rates for infection by malaria parasites were higher in the dizygotic than the monozygotic twin pairs whereas concordant rates for fever were significantly higher among the identical twins. This study provided evidence that host genetic factors affect risk of malaria disease rather than infection. On the other hand, the development of fever is influenced by host genetic factors. These findings were in agreement with the hypothesis that monozygotic twins could share alleles involved in the production of pyrogenic cytokines (e.g. TNF-α). High heritability for proliferative responses to several malaria antigens was demonstrated in a study with 267 Gambian twin pairs, including 60 monozygous twins (Jepson, Sisay-Joof et al. 1997).

A large number of host genes that confer differential predisposition to various manifestations of malaria have been identified with association analysis (reviewed in (Frodsham and Hill 2004)). Unfortunately, the discovery of these polymorphisms has not led to the development of new treatments or prophylaxis against malaria. Most of these mutations carry serious consequences for the host when homozygous and therefore have been offered no practical therapeutic lead.

The malaria heritability study by Mackinnon and colleagues, and the familial segregation analysis of immunological responses to malaria antigens in Papua New Guinea indicated that Mendelian effects might govern certain antigen responses but the overall picture is complex

(Stirnadel, Beck et al. 1999; Mackinnon, Mwangi et al. 2005). It is more probable that the genetic basis of susceptibility to malaria is mostly the result of many different polymorphisms with relatively modest effects on malaria disease outcome.

1.12. Malaria is a strong selective pressure shaping the human genome

Natural selection is the process whereby some of the inherited genetic variation between individuals will result in differences in their ability to survive and reproduce successfully. Haldane first pointed out that an infectious disease causing high mortality among children could be important in shaping human evolution by exerting selective pressure on mutant genes protecting against that infection (Haldane 1949). Evidence is steadily accumulating that *falciparum* malaria fits Haldane's description of such an infectious disease.

Recent studies have shown that though the exposure of human populations to malaria is relatively short, it has left its mark as a wide range of genetic diversity (Weatherall, Miller et al. 2002). It is thought that severe malaria became endemic ~10,000 years ago coincident with the origination and spread of agriculture in the Middle East and Africa (Tishkoff, Varkonyi et al. 2001; Joy, Feng et al. 2003), but mild forms of the disease may have existed in humans throughout much of their evolutionary history. In malaria endemic regions, most children present with the mild form of the disease. Only a small percentage of those infected go on to develop severe or complicated disease and consequently die of it. This is mainly explained by host resistance factors that have evolved over several thousand years of selection under the pressure of high exposure to *falciparum* malaria (Miller, Good et al. 1994).

The high frequency of otherwise deleterious mutations in areas where malaria is endemic is probably due to the protection that they confer against severe malaria. It is assumed that

mortality from malaria is the sole selective force for those mutations (Kwiatkowski 2005). The classical malaria resistance genes are the best examples of this natural selection. Genetic polymorphisms of the innate immune system and of human erythrocytes have thus been proposed as factors protecting against severe malaria (Wilkinson and Pasvol 1997; Flint, Harding et al. 1998; Cooke, Mohandas et al. 2004) (Roberts and Williams 2003). The protective effect of certain red blood cell polymorphisms (Haemoglobin S [HbS], HbC, HbE, α-thalassemia, β-thalassemia, and ovalocytosis) against severe *falciparum* malaria is well-known (Agarwal, Guindo et al. 2000; Modiano, Luoni et al. 2001; Weatherall, Miller et al. 2002; Cooke, Mohandas et al. 2004).

It is important to identify host determinants of disease susceptibility and protective immunity within African populations. The host defence mechanism against malaria is complex (Frodsham and Hill 2004). New strategies are required to identify the most important targets of protective immunity in malaria. Analysis of positive selection signals like the differential distribution of unusually extended haplotypes between cases and controls could identify the loci under most intense selection. Given that malaria exerts the most powerful selective force on the human genome that we know, the ability to detect genetic determinants of susceptibility to malaria should be more feasible than in other complex diseases.

1.13. Utilizing ethnicity to tackle the question of malaria

Geneticists are interested in finding genes associated with disease. Because of widespread health disparities, ethnicity is a variable that is often said to be relevant in this context. The idea is that members of a preconceived ethnic group share common ancestry that may include genetic risk factors. Human variation has been shaped by the long-term processes of

population history, and population samples that reflect that history carry information about shared genetic variation or ancestry.

Differences in susceptibility to malaria between ethnic groups have been observed in many studies (Greenwood, Groenendaal et al. 1987; Terrenato, Shrestha et al. 1988). Studies carried out in both Nigeria and in the Gambia have indicated that the frequency of splenomegaly is higher in people belonging to the Fula ethnic group than in other sympatric groups (groups inhabiting the same geographic region) (Bryceson, Fleming et al. 1976; Greenwood, Groenendaal et al. 1987).

Although it is often difficult to be confident about ascribing a genetic cause to apparent differences between populations, studies on susceptibility to malaria in different sympatric ethnic groups in Burkina Faso (Mossi, Rumaibe, and Fula) (Modiano, Petrarca et al. 1996; Luoni, Verra et al. 2001; Modiano, Luoni et al. 2001) and in Mali (Fula and Dogon) (Dolo, Modiano et al. 2005), are of particular interest. Genetic factors appear to underlie the striking differences in resistance to malaria between these groups. The Fulani group, show a stronger immune response to malaria antigens and greater resistance to both parasitisation and disease than the other two ethnic groups resident in the same area. Environmental factors and known malaria resistance alleles did not account for these differences. The Fulani have lower frequencies of known protective variants such as HbS. This fact suggests that immune response genes yet to be identified were responsible. In a recent study, parasitological and immunological parameters were compared between Fulani and the neighbouring ethnic group, the Dogon of Mali. The study was performed outside the malaria transmission season and all individuals included were healthy at the time. The study confirmed that the Fulani were less parasitized, had fewer circulating parasite clones in their blood and had significantly higher levels of antimalarial IgG (IgG1 and IgG3) and IgE antibodies compared to the Dogon (Bolad, Farouk et al. 2005).

Chapter 2:

Study Samples, Materials, and Methods

2.1. Sampled populations and study area

DNA samples used in the various studies in this thesis were from three sources.

I. With the help of members of the Institute of Endemic Diseases, University of Khartoum, I collected population samples, mostly family trios, from two villages along the eastern bank of Rahad river area of eastern Sudan along the Sudanese-Ethiopian border.

II. Joining efforts with other members of The Kwiatkowski group at the Wellcome Trust Centre for Human Genetics (WTCHG) in Oxford, I was able to use the second set of samples which was the HapMap Yoruba and CEPH collection obtained as EBV-immortalized lymphoblastoid B-cell lines from the Corriell Repository (Coriell Institute for Medical research).

III. The third data set was that obtained from MalariaGEN whole-genome association study in Gambian samples. An endeavour which involved the collaborative efforts of a huge number of people, in the Gambia (M. Jallow, M. Pinder and colleagues), WTCHG in Oxford, and the Wellcome Trust Sanger Institute in Cambridge, UK.

The following few sections of this chapter will consist of detailed descriptions of these three datasets.

2.1.1. Sudanese samples

The Sudanese samples of my study came from two populations of recent migrants inhabiting two neighbouring villages in an area endemic to malaria in Eastern Sudan. Koka village was established around 1940 by Hausa people originating from West Africa. Salala village was established around 1960 by Masalit people originating from western Sudan.

The two villages lie along the eastern bank of Rahad river area of eastern Sudan along the Sudanese-Ethiopian border, about 400 kilometres south-east of Khartoum and about 200 kilometres south-west of Gedarif town (Figure 2.1.1.), in the largest mechanized agricultural area in Sudan. Koka village is 35 km north of Salala village.

Figure 2.1.1. A map of Sudan.
The circled area designates the region in eastern Sudan where the study was conducted.

2.1.1.1. Background on the Sudanese populations and their demography and environment

The Hausa:

The Hausa population in Africa which is estimated at 22,000,000 makes up the largest ethnic group in all of West Africa. It is a farming population whose social organization is based on a hierarchical system based on occupation, wealth, and birth. Massive migration to Sudan occurred during early 20th century, motivated by the interest of the British administration in the Hausa as skilled farmers. The Hausa language is part of the Chadic branch of the Afro-Asiatic family of languages.

Koka village was established about 60 years ago by the emigrant Hausa tribe from northern Nigeria (mainly from the towns of Kanu and Sakatu). The Koka community is a closed community, male polygamy highly practiced. Above 90% of all marriages and births in the past 40 years occurred within the village. Only 1% of its inhabitants were born in Nigeria and migrated from there, the rest were either born in the village or incorporated to it through marriage from nearby Hausa villages. Village size in 2004 was 1521 individuals, making around 67 large extended families with 30 individuals per family on average.

The Masalit people:

Of the peoples living in the western province of Darfur who spoke Nilo-Saharan languages and were at least nominally Muslim, the most important were the Masalit. They were primarily cultivators living in permanent villages, but they practiced animal husbandry in varying degrees. The Masalit, living on the Sudan-Chad border, were the largest group.

Um Salala village is located in the eastern bank of the River Rahad (Gadaref State, Sudan). The village was founded between the years 1969 and 1984. It was founded by members of the Masalit tribe who migrated from El-Geneina in Darfur state, western Sudan and settled along the Rahad River. The migration to the village increased dramatically after the drought

that hit Darfur in 1984. In Salala village, 50% of the population were born in the village, the others were born in western Sudan and migrated to this area in the east, either as part of the first migration that founded the village or as part of smaller migrations since that time. About 28% of marriages among the Masalit were within the village. According to a census carried out in 2004, the estimated village size was 1309 individuals. Most Masalit live as nuclear families in the village settlement, and there are about 300 small families in the village with 5 individuals per family on average. The Masalit language is part of the Nilo-Saharan family of languages.

Polygamy is common in both villages but consanguinity is rather occasional.

Population growth:

Good population data is lacking for my study area in eastern Sudan, because it is one of the least accessible and hence least developed in the country. Medical services are scarce or nonexistent due to its remoteness. Nevertheless, inference about population growth can be made from the country-wide trend. In 1990, the National Population Committee and the Department of Statistics put Sudan's birth rate at 50 births per 1,000 and the death rate at 19 per 1,000, for a rate of increase of 31 per 1,000 or 3.1 percent per year. This is a staggering increase; compared with the world average of 1.8 percent per year and the average for developing countries of 2.1 percent per annum, this percentage made Sudan one of the world's fastest growing countries.

Environment:

From January to March, there is practically no rainfall countrywide. By early April, the moist southwesterlies reach southern Sudan, bringing heavy rains and thunderstorms. By May the moist air reaches eastern Sudan which has a typical savannah climate with a short rainy season (May-October) and a dry hot one (November -May).

In August the rains in the Ethiopian highlands swell the Blue Nile until it accounts for 90 percent of the Nile's total flow. The Blue Nile's two main tributaries, the Dinder and the Rahad, have headwaters in the Ethiopian highlands and discharge water into the Blue Nile only during the summer high-water season. For the remainder of the year, their flow is reduced to pools in their sandy riverbeds.

The ecology of the area around Koka village is the same as in Salala village. The soil in the area is alluvial clay that forms large cracks during the dry season. The area is characterized by woodland dominated by *Balanites aegyptiaca* and *Acacia seyal* trees. *Balanites/Acacia* woodland, the cracked clay soil and termite hills are the typical biotope of Phlebotomus orientalis, the vector of visceral leishmaniasis (kala-azar) in the Sudan.

2.1.1.2. Age distribution in the two Sudanese villages

In a study of the epidemiology and clinical spectrum of *L. donovani* infection, that took place in the two villages of my study, from April 1994 to April 1996; the age and sex distribution in the two villages is shown in Figure 2.1.1.2. The age distribution is typical of that of developing countries, with a high proportion of younger age groups. There was a significant difference in sex ratio with apparent under representation of young males in Salala, likely because this group work as casual labourers outside the village (Khalil, Zijlstra et al. 2002).

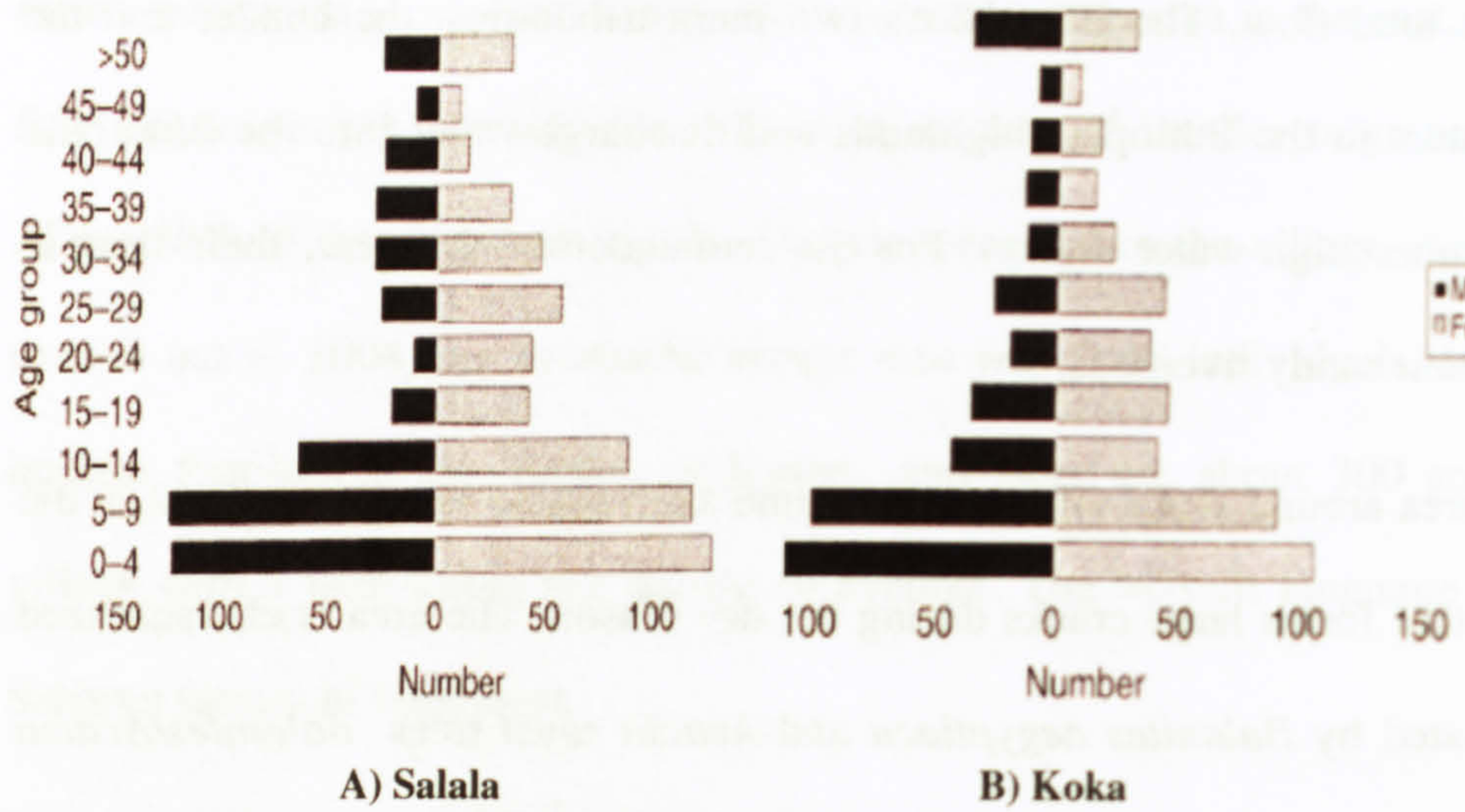


Figure 2.1.1.2: Population structure in A) Salala and B) Koka villages.
Horizontal bars represent number of males and females in each age group. (Khalil, Zijlstra et al. 2002).

2.1.1.3. Socio-economic and nutritional parameters

In Salala village, the inhabitants are mainly farm labourers and subsistence farmers and live in grass huts without latrines. The standard of hygiene is poor. Their diet is mainly carbohydrates with meat and fish once or twice a week; lemons and mangoes are available during the season.

The Hausa in general have a better standard of living and nutritional status. They also have a more structured hierarchy in the social order. Houses in the village are thatched grass huts and all have latrines. The village appears much cleaner than Salala. The villagers' diet consists mainly of fish supplemented with sorghum, millet, vegetables and fruit.

To compare the nutritional condition in the two study villages, for all children under 5, Khalil, E.A.G. and colleagues, 2002, calculated Z-scores for weight-for-age (WAZ), height-for-age (HAZ) and weight-for-height (WHZ), using the EPINUT anthropometry program of EpiInfo; they then compared the means of these parameters using the t-test. For adults, the

body mass indexes (BMI) were calculated, and mean values were compared between the two villages (Table 2.1.1.3.).

Both mean WAZ and WHZ scores were lower in Salala, indicating poorer nutritional state. Also in adults, mean BMI values in Salala were significantly lower. Splenomegaly rates were higher in Koka, in those over 15 years of age.

	MK	Um-Salala	*P*-value
Z scores			
WAZ	−1.27 (1.47)	−1.88 (1.2)	0.000
HAZ	−2.04 (1.92)	−2.17 (1.7)	0.40
WHZ	+0.09 (1.30)	−0.68 (1.12)	0.000
BMI	22.3 (9.0)	20.1 (2.8)	0.000
Spleen size (age ≤ 15)			
Mean spleen size in cm (SD)	1.37 (2.19)	0.49 (1.59)	0.000
0	329 (64%)	570 (87)	0.000
1–4	129 (25%)	54 (8%)	0.000
5–9	58 (11%)	30 (4.5%)	0.000
10–14	2 (0.4%)	2 (0.3%)	1.000
≥15	0	1 (0.2%)	1.000
Spleen size (age > 15)			
Mean spleen size in cm (SD)	0.47 (1.67)	0.19 (1.68)	0.004
0	327 (91%)	428 (96%)	0.15
1–4	17 (5%)	10 (2%)	0.07
5–9	16 (4%)	7 (1.5%)	0.02
10–14	–	2 (0.5%)	0.50
≥15	1 (0.3%)	0	0.44
Malaria prevalence*			
April 94	498/694 (72%)	161/790 (20%)	0.000
November 94	170/222 (77%)	118/255 (46%)	0.000
April 95	180/227 (79%)	218/264 (83%)	0.42
November 95	75/119 (63%)	79/305 (26%)	0.000

WAZ: Weight-for-age Z-score, HAZ: height-for-age Z-score, WHZ: weight-for-height Z-score.
* In those eligible for testing: fever, ill or splenomegaly.

Table 2.1.1.3: Nutritional parameters, spleen size and malaria prevalence in Koka (MK) and Salala (Um-salala). (Khalil, Zijlstra et al. 2002).

2.1.1.4. Disease epidemiology at the study site in Eastern Sudan

The two populations of Hausa and Masalit moved to the area of the Rahad River within the past 50 years. The area used to be a game reserve. Visceral leishmaniasis and malaria were

both endemic as well as other infectious and water-borne diseases, making the settlement a challenging and costly undertaking in terms of health. The Masalit in particular succumbed to visceral leishmaniasis.

This area is the major endemic area for visceral leishmaniasis in Sudan (75% of reported cases in 1987). Khalil, E.A. et al. (2002), found there are differences between the two populations in their susceptibility to leishmania. In 1996 the disease rate was 38.3/1000 person-years in Salala, with 1.2:1 ratio of clinical to subclinical infection. In Koka, it was 4.6/1000 person-years, and there was a 1:11 clinical to subclinical infection ratio.

The scale of the malaria problem in these villages has received relatively little attention in the past because the area is extremely difficult to access by road during the rainy season. But previous studies of VL provide data for the dry season, when there appears to be a significant difference in malaria prevalence between the villages.

The area is endemic to malaria with unstable transmission in a meso-endemic to hyper-endemic levels depending on the season. Most infections are due to *P. falciparum* with a minority caused by *P. malariae*.

In October 2002, the final month of the rainy season, the diagnostic laboratory of the local health centre, at a nearby village, performed 1200 blood films on patients presenting with fever, of which half were found to have *P. falciparum* parasitaemia. Since the population of the group of villages covered by the health centre is around 15,000, this observation suggests that at least 4% of the population, averaged across all age groups, had malaria fever episodes during that single month. This is a minimum estimate and the true value may be much higher than that. In surveys carried out shortly after the end of the rainy season, spleen rate was found to be around 25-30% in children below 15 years-old.

Data for malaria prevalence in the two villages were obtained from eight cross-sectional surveys between 1994 and 2006, which were carried out by members of the Institute of Endemic Disease, University of Khartoum (http://www.iend.org/). All cross-sectional surveys prior to 2004 showed little or no clinical malaria in the two villages. From 2004 and onwards, the cross-sectional surveys were supplemented by an active surveillance system involving village workers and teams that stayed over during the rainy season. Longitudinal follow up of these two populations have indicated that despite the malaria endemicity, the disease is generally presented in mild clinical forms particularly in Hausa. Their recent West African origin and agricultural life style suggests an early and stable exposure to the malaria parasite, with extended periods of co-evolution. Clinical malaria was consistently few fold higher in Masalit than Hausa.

Clinical malaria was mainly manifested with parasitemia and a body fever of 38-42 ˚C. Vomiting, headache and other symptoms were occasionally reported. Severe malaria as defined by WHO criteria for malaria was rarely encountered. Severe anaemia was found to be relatively infrequent in the two populations with a mean haemoglobin level in the Hausa population of 14.2± 1.08 g/dl and for the Masalit 13.1±1.78 g/dl.

2.1.1.5. Sample recruitment

The samples I used for my study were collected during a cross sectional survey conducted during June 2004. The study was reviewed and ethically approved by the Ethical Committee of the Institute of Endemic Diseases, University of Khartoum. Samples were taken with informed consent from all individuals (see appendix for informed consent documents).

The criteria I chose for sampling trios (two parents and a child) from the two villages, in order to better represent the general village population; is that the parents had to be

unrelated, neither to each other within a trio, nor to parents from other trios included in the study. The parents in each family that was seen were questioned on their degree of kinship, and they were only sampled if they were found not to be related. Then, one of their children was chosen at random to be sampled. To make sure that there were no background relatedness between the chosen trios, I had to construct the pedigree for the whole village using the software Cyrillic (Figure 2.1.1.5).

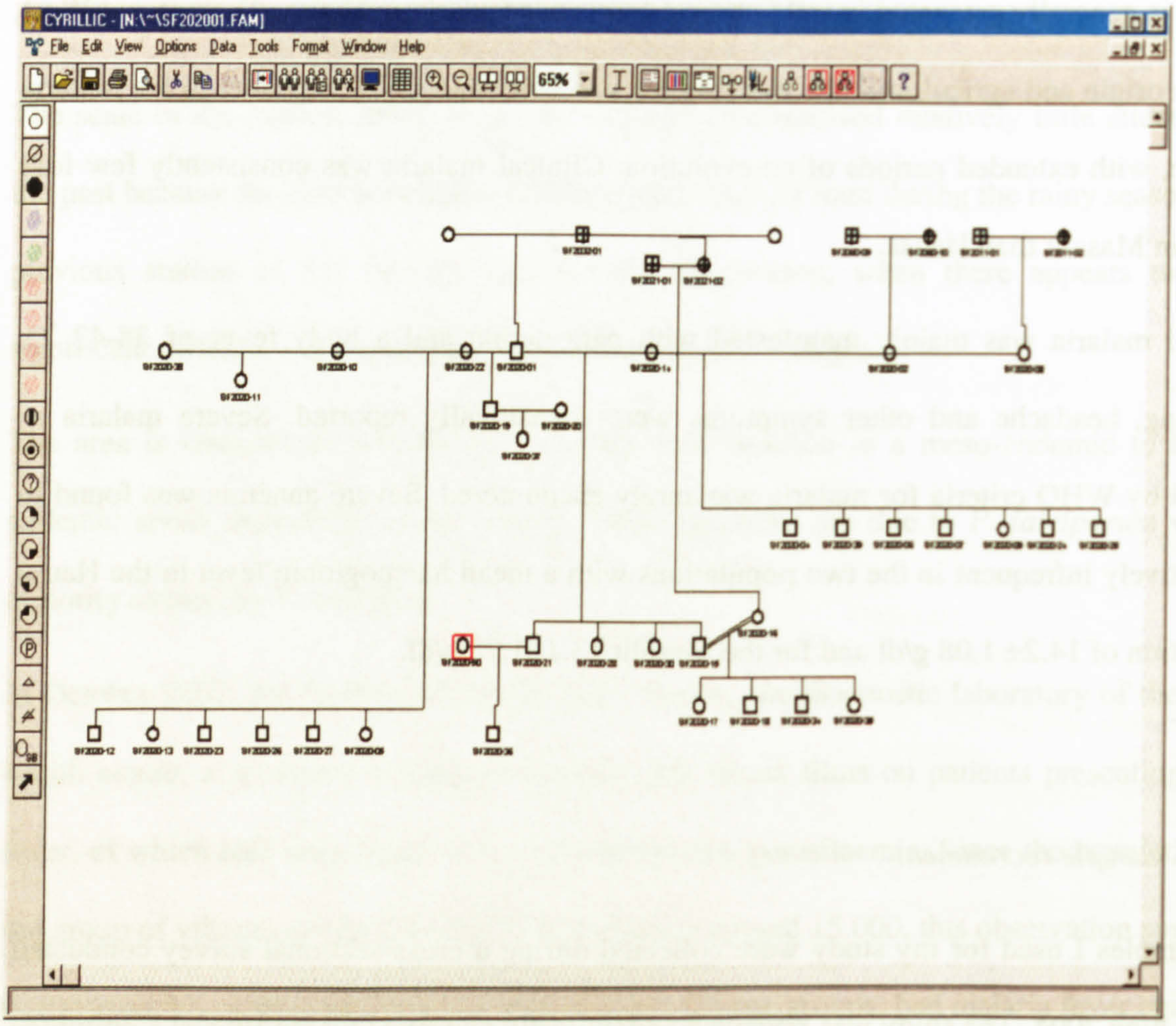


Figure2.1.1.5: A window in Cyrillic showing an example of an average sized family in Koka village.
The red box indicates the individual for whom the family tree is retrieved.

Firstly, each family had its family tree constructed separately. Then, families were connected together through mutual individuals by reviewing family histories for consanguinity and

presence of relatives in the village. Only those trios with no kinship either within or between them were chosen for the study.

This task proved to be very difficult due to the many connections between families, so much so that the whole village resembles one big pedigree. Because of the requirement to choose trios who are not related, there might have been a disproportionate representation of individuals originally from outside the village and have only been incorporated into it by marriage.

Data collection questionnaire for the Sudanese samples

ID:.....House number:.....

Date:.....

Village:..... Tribe:.....

Name:.....

Mother's name:.....

Date of Birth (age):.....Sex:.....

Are(you/ your mother & father) related? .....

If yes what is the degree? .....

Do you have relatives in the village?.....
If yes ,
name.....& D of relatedness.....

name.....D of relatedness.....

occupation:.....

Residence in the last 6 months.....

Do you use bed-nets?.....
If yes , how often?.....

Past Medical History:

Did you have Malaria before?.....

If yes,How frequent do you get it(/year)?.....

how long ago was the last attack?.....

methods of diagnosis .....

symptoms of the last attack.....

what treatment did you take for it .....

what was the response to treatment?.....

did you ever had severe malaria before?.....

did you ever had kala-azar before?.....

if yes how long ago?..... who diagnosed it?.....

did you have PKDL before?.....

Present complaints:

Fever ☐

if yes, duration.....grade.....type.....

sweating ☐	Rigors ☐	Diarrhea ☐
Nausea ☐	Vomiting ☐	abdominal pain ☐
Headache ☐	drowsiness ☐	Convulsions ☐
Cough ☐	Generalweakness ☐	Allergy ☐

Do you have any other complaints?.....

Examination:

General appearance.....

Temperature.....

Spleen (cm).....

PKDL.....

Other examination.....

Investigations:

BF.....

PCV.....

Diagnosis and Treatment:

.....

2.1.2. The HapMap samples

The DNA samples for the HapMap have, so far, come from a total of 270 people. The Yoruba people of Ibadan, Nigeria, provided 30 sets of samples from two parents and an adult child. In Japan, 45 unrelated individuals from the Tokyo area provided samples. In China, 45 unrelated individuals from Beijing provided samples. Thirty U.S. trios provided samples, which were collected in 1980 from U.S. residents with Northern and Western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH).

The non-profit Coriell Institute for Medical Research provides DNA and cell lines from the samples for research projects that have been approved by the appropriate ethics committees. The samples and cell lines are not linked to any individual in the populations studied. However, the samples and cell lines are identified as coming from one of the four populations participating in the study.

The HapMap samples were collected with the understanding that they would be used not only to develop the HapMap, but also for a wide range of future studies. Their use is not limited to the study of any particular disease. Each community that provided new samples for the HapMap has established a Community Advisory Group (CAG) to serve as a liaison between the community and the repository. The CAGs receive quarterly reports that list the investigators who requested their samples and the nature of the research those investigators plan to conduct with their samples. They also receive a quarterly listing of the major publications that result from any research with their samples, so that they can stay apprised of how their samples are being used. The entire donor community, through its CAG, could decide to withdraw its samples from the Repository if it were determined that community's samples were being used in a manner inconsistent with the wishes of most of the members of that community. According communities a right to withdraw their samples in this manner is

consistent with contemporary standards of research ethics for genetic variation research that involves identified populations (HapMap 2004).

In this thesis, I genotyped and used publicly available data for two sets of samples genotyped in phase 1 and 2 of the international HapMap project:

Yoruba in Ibadan, Nigeria (abbreviation: YRI): A set of 30 Trios with 90 samples. These samples were collected in a particular community in Ibadan, Nigeria, from individuals who identified themselves as having four Yoruba grandparents. The sample set does not necessarily represent all Yoruba people, whose population history is complex.

CEPH (Utah residents with ancestry from northern and Western Europe) (abbreviation: CEU): A set of 30 Trios with 90 samples. These samples were collected from people living in Utah with ancestry from Northern and Western Europe. The term "CEPH" stands for the Centre d'Etude du Polymorphisme Humain, the organization that collected these samples in 1980. Because the importance of precision in assigning group membership to prospective donors based on ancestral geography was not well appreciated in 1980, it is unclear how accurately these samples reflect the patterns of genetic variation in people with Northern and Western European ancestry.

2.1.3. The Gambian samples

Study site

The Gambia can be crudely divided into coastal regions (know as Kombos), with a predominantly urban population, and inland regions (non-Kombos) largely rural and relatively less developed. In the Gambia malaria transmission is highly seasonal and it usually clusters between July and December, coinciding with the arrival of the rains. The Royal Victoria Teaching Hospital (RVTH), located in the capital Banjul on the Atlantic

coast, is the main referral hospital in the country and an average of 300 children are annually treated for severe malaria.

Data collection

Since 1997 a research team from the Medical Research Council (MRC) in The Gambia has actively been recruiting patients presenting to the RVTH paediatric ward with signs of severe malaria. On admission, a detailed clinical history, physical examination including assessment of consciousness (Molyneux, Taylor et al. 1989), capillary blood samples for haematocrit and thick blood smears for malaria parasites were conducted on all children. The study team reviewed the children within 8 hours of admission. If consent was given, children with severe malaria were recruited if they were aged 4 months to 14 years, with *P.falciparum* malaria and fulfilled one or more of the WHO criteria for severe malaria (2000). For anaemic children requiring transfusion, recruitment was done at the time of blood transfusion. Demographic details including ethnicity, place of residence, treatment received before presentation and clinical details were recorded on to standard forms. A four-hourly evaluation of the temperature, pulse, respiration, Blantyre Coma Score (which was designed by Drs Terrie Taylor and Malcolm Molyneux in 1987 as modification of the Pediatric Glasgow Coma Scale. The score is a number from 0 to 5, determined by adding the results from three groups: Motor response, verbal response, and eye movement. The minimum score is 0 which indicates poor results while the maximum is 5 indicating good results), and blood sugar estimation were performed. Quantification of urine output and blood pressure were done only in children with clinical suspicion of shock or acute renal insufficiency. Thick blood films were examined after Field's staining and read independently by a second experienced microscopist. A slide was classified as negative if no parasites were seen in 100 fields. The parasite concentration was estimated by counting the number of parasites per high power field and multiplying by 500; one parasite per high

power field was taken to indicate parasitaemia of approximately 500 per μL (Greenwood and Armstrong 1991). A final reading was made by a third microscopist only if there were large discrepancies. Blood glucose levels were measured using BM-Test-Glycaemic 1-44 strips. The study was approved by the Gambia Government/Medical Research Council (MRC) Ethics Committee. Written informed consent was obtained from the patient or family after discussion in local language.

Collected samples were used to establish several large population and family-based case-control association studies. For family-based association analysis, both parents of affected children served as control subjects.

Clinical case definitions of severe malaria and inclusion criteria

Children were categorized as having cerebral malaria if they had a Blantyre coma score ≤3, at least 30 minutes after the last seizure and appropriate treatment of hypoglycaemia, with asexual forms of *P. falciparum* on blood film and no other evident cause of coma (Molyneux, Taylor et al. 1989). Severe malarial anaemia was defined as packed cell volume ≤15% or haemoglobin ≤5 g/dl with asexual forms of *P. falciparum* on blood film. Respiratory distress was defined as the presence of any of the following: deep breathing, abnormal respiratory pattern, grunting and/or use of accessory muscles of respiration.

Only cases with complete information for ethnicity and area of residence were selected. Moreover, cases with unknown coma score or haemoglobin concentration were not included. Priority was given to individual cases living in the Kombos over rural Gambia.

2.2. DNA collection and extraction

2.2.1. Sudanese samples DNA collection and extraction

With the help of other members of the Institute of Endemic Diseases, I collected DNA samples using the buccal brush method, and extracted them by the guanidine hydrochloride method as follows:

- 20uL of 10mg/mL proteinaseK was added to each sample to rupture cell membranes and degrade protein.
- 1mL of 6M guanidine chloride and 300 µL of 7.5M NH_4 acetate were added to each sample in a 50 mL tube and left overnight to remove remaining proteins and peptides.
- Samples were centrifuged at 2000 rpm and the supernatant transferred to 15 mL polypropylene tube containing 2 mL of pre-chilled chloroform.
- Those were vortexed, left to stand for one minute and centrifuged for 5 minutes at 2500 rpm.
- The upper layer was collected and transferred to a 15 mL Falcon tube containing 10 mL of cold absolute ethanol; the tube was then inverted and shaken gently.
- Tubes were kept at -20°C for at least 2 hours then centrifuged at 3000 rpm for 15 minutes.
- The supernatant was carefully drained and the pellet of DNA was washed with 4 mL of 70% ethanol.
- Tubes were centrifuged at 3000 rpm for 15 minutes, after which the supernatant was carefully drained. The tubes were left inverted on a paper towel for 5 minutes.

- The pellet was washed again using 70% ethanol, the supernatant was discarded and the pellet was left to air-dry.
- Pellets were re-suspended in 100-200 µL of de-ionized water and left for 2 days at 4 ºC before being quantified.

2.2.2. HapMap samples DNA extraction

EBV-immortalized lymphoblastoid B-cell lines from the HapMap Yoruba (Ibadan, Nigeria) collection were obtained from the Corriell Repository (Coriell Institute for Medical Research). Other members of the Kwiatkowski lab were responsible for growing the cell lines in cell culture at the Oxford WTCHG (Suzana Campino, Sarah Auburn and colleagues). Cell lines were maintained at between 200,000 and 800,000 cells per mL in RPMI 1640 medium (Sigma) with 10% fetal calf serum (Sigma), 200 mM L-glutamine, penicillin, and streptomycin (all Sigma) at 37°C in humidified incubators, in an atmosphere of 5% CO_2. DNA was extracted from aliquots with 20 million cells.

gDNA extraction from lymphoblastoid B-cells was undertaken by myself and Anna Richardson (Kwiatkowski lab manager) using the Nucleon kit from Tepnel Life Sciences (http://www.tepnel.com/) according to their protocol.

2.2.3. Gambian samples collection and DNA extraction

In Gambia local teams extracted DNA samples from venous blood samples (~1mL) using the Nucleon BACC2 DNA extraction Kit, as per protocol. Aliquots of the extracted DNA samples were sent to the WTCHG, Oxford, UK.

DNA samples were whole genome amplified using ϕ29 multiple displacement amplification (MDA) with REPLI-g™ 625S reagents based on instructions from the manufacturer (MSI Inc, New Haven). The quality and quantity of DNA was assayed in each sample prior to amplification with PicoGreen. The DNA samples were in TE (10mM Tris-HCl pH 7.5, 1mM EDTA) and the concentration was ≥ 20ng/μL in a total volume of ≥ 10μL. Amplified DNA samples were re-assayed using PicoGreen, normalized to 250 ng/μL, and loci representation Quality Control (QC) was performed by Taqman assay on two loci. All ϕ29MDA DNAs selected for genome-wide genotyping resulted from reactions with a minimum of 5 ng input genomic DNA. The quality of the amplified DNA was further assessed by assaying 30 SNPs across the genome using Sequenom iPlex.

2.2.4. Sample Archiving

Collaborative efforts have enabled the collection of thousands of DNA samples for the genetic epidemiologic analysis of severe malaria in the Kwiatkowski laboratory. Thus efficient database systems were required for managing large numbers of DNA samples and their associated clinical information. I carried out the archiving of the Sudanese set of DNA samples, and helped with the archiving of the HapMap cell lines.

Samples were labelled with a new ID for confidentiality. The DNA concentration of each sample was measured using the PicoGreen method and the results were recorded in the DNA archive database and used to prepare a diluted set of samples to a standard DNA concentration. The samples were all diluted with 1x TE to a standard concentration of 20ng/μL. Five microlitres of this dilution were transferred to a deep-well plate and diluted to 1ng/μL. The samples were stored at -20°C at dilutions of 20ng/μL in screw-top tubes and at

1ng/μL in deep well plates. The original sample was returned to -80°C. The exact location of each sample (box position, box location and freezer) was recorded in the DNA archive.

A plate plan containing the position of each sample was created and a duplicate copy kept in the group's secure database.

2.2.5. DNA quantification: PicoGreen

At the WTCHG in Oxford, DNA samples were quantified and archived. DNA samples were prepared to give 20ng/μL stock samples. I quantified the sample DNA concentrations using the PicoGreen® assay (Molecular Probes, Leiden, Netherlands). PicoGreen binds to double stranded DNA and fluoresces under UV light in proportion to the amount of DNA bound.

Preparations before sample reading

The first two stages of the PicoGreen protocol were the preparation of TE and the preparation of DNA standards. From the 20x TE provided with the PicoGreen kit, 1 mL was diluted with 19mL of distilled water to give 1x TE. Stock DNA of 100μg/mL lambda DNA was provided with the PicoGreen kit. From this stock, 54μL were added to 2646μL TE to give 2700μL of DNA at 2μg/mL. DNA samples were diluted by taking 2μL of each sample and adding 200μL of the diluted TE. Stock PicoGreen from the kit (25μL) was added to 4975μL TE to give 5mL working reagent, sufficient for one 96-well plate of samples. The light-sensitive working reagent was stored in a 15mL falcon tube wrapped in tin foil at 4°C and used within half an hour of being made.

Sample Reading

From each DNA standard (in duplicate) and each DNA sample (in duplicate), 50μL were added to a Costar 3925, black polystyrene assay plate (Figure 2.2.5.). PicoGreen reagent

(50μL) was added to each well. The solution was mixed and incubated for 5 minutes at room temperature in darkness. Fluorescence was measured using the SPECTRAfluor Plus fluorimeter (Tecan instruments, Reading, UK). The DNA concentration was calculated using a standard curve of fluorescence versus DNA concentration. The integrity of the standards was judged by their fit to a standard curve. An R^2 value of > 0.9 was accepted. Where several discrepancies occurred, the measurements were redone using different standards on another fluorimeter (Cytofluor). The sample concentrations were entered into the sample archive database.

	1	2	3	4	5	6	7	8	9	10	11	12
A	S	S	1	1	9	9	17	17	25	25	33	33
B	T	T	2	2	10	10	18	18	26	26	34	34
C	A	A	3	3	11	11	19	19	27	27	35	35
D	N	N	4	4	12	12	20	20	28	28	36	36
E	D	D	5	5	13	13	21	21	29	29	37	37
F	A	A	6	6	14	14	22	22	30	30	38	38
G	R	R	7	7	15	15	23	23	31	31	39	39
H	D	D	8	8	16	16	24	24	32	32	40	40

Figure 2.2.5: PicoGreen® Assay: Plate Plan.

2.2.6. Whole-genome Amplification

Whole-genome amplification was required to increase the number of genome copies of the DNA samples. Samples were subject to whole-genome amplification using the PEP (Primer Extension Preamplification) protocol (Zhang, Cui et al. 1992). The PEP PCR reaction uses a number of 15 nucleotide, random sequence primers called "N15" primers. These primers anneal to various positions on the template DNA wherever they find their complementary sequence and, thus, amplify the entire genome.

Five microlitres of the stock DNA samples at 1ng/μL was added to the corresponding well on a skirted Thermofast 96-well plate. PEP master mix was prepared with the following volumes of reagents per sample: 5μL PCR buffer, 2.5μL $MgCl_2$ (50mM), 2μL of dNTP mix (each dNTP at 5mM), 2.2μL N15 primer mix (all GENPAK Ltd), 32.8μL MilliQ water and 0.5μL Biotaq (Bioline). Thus, 45μL master mix was added to each well. The cycling conditions were: 94°C for 3 minutes followed by 50 cycles of 94°C for 1 minute, 37°C for 2 minutes, ramp at 0.1°C per second to 55°C, 55°C for 4 minutes and a final extension period at 72°C for 5 minutes. For quality control, 2 μL of random samples of the PEP product was run on a 2% agarose gel (Sigma, USA).

2.3. Choice of markers

In this thesis, different strategies were used for SNP selection in different candidate regions. The objective was to achieve relatively uniform marker spacing across the genomic regions of interest. Following the initial selection of a set of SNPs by any strategy, the final set on which analysis will be carried out is likely to differ as various SNPs may be eliminated on account of assay design failures and economic multiplexing decisions.

From the abundance of available SNPs, inappropriate markers may be filtered out by exclusion of non-validated SNPs, with preference for validated SNPs especially in an African population. A minimum reasonable frequency should be required for marker inclusion. Indeed, from a public health perspective it is important to note that rare alleles in common disease generally explain relatively little of the overall disease prevalence (low population-attributable fraction, or PAF). The Common Disease Common Variant

hypothesis proposes that common modest-risk alleles may account for a greater PAF in common disease than do rare high-risk alleles.

LD, the non-random association between alleles at different loci, is an important consideration in marker selection, particularly in large genetic regions where it is not economically feasible to genotype a high density of SNPs. In chromosomal regions where there is substantial LD between markers, allelic dependence improves the chances of establishing the approximate location of a disease mutation without actually typing it. When a mutation arises in a population, it will be on a specific haplotype. Once the haplotype structure is determined, redundant SNPs can be identified and it is possible to type the minimum number of SNPs required to uniquely tag all the haplotypes to search for a mutation (Daly, Rioux et al. 2001; Gabriel, Schaffner et al. 2002). The use of haplotype tagging SNPs (htSNPs) enables researchers to capture the majority of the haplotypic variation and reduce the amount of genotyping required to scan a genetic region.

2.3.1. Marker choice in 5q31 genomic region

Markers typed in the 5q31 region were chosen from a larger set that had previously been tested in the laboratory at the WTCHG in Oxford in samples from The Gambia and the UK. There were 34 markers selected as the most efficient set of markers to capture most of the haplotypic diversity in those populations (haplotype tagging SNPs).

Previously, for two population samples from the Gambia and the UK, initially, 162 SNPs were identified from dbSNP (ncbi.nlm.nih.gov/SNP) and the literature. Selected SNPs were those reported to have MAF >0.05, and that together gave a good representation of coding and noncoding regions. A total of 98 SNPs were analysed after excluding those that were out

of Hardy–Weinberg equilibrium, had a genotyping failure rate >10%, or had MAF <0.05 in both populations tested. The West African sample comprised 32 mother–father–child trios where the child had severe malaria. In order to identify haplotype tagging SNPs in the 5q31 region, firstly, haplotypes of the 128 parental chromosomes from each population were determined from the pedigree data using the **PHASE** algorithm, then an unstructured approach using ENTROPY which determines the information content of each SNP without consideration of block structure was used for selecting tagging SNPs. This method identified 21 tagging SNPs in Europeans and 18 in West Africans (Luoni, Forton et al. 2005).

The marker set I subsequently typed in the Sudanese populations consisted of those 18 Gambian tagging SNPs, in addition to 16 others identified as having a long-range LD pattern in the region (Sadighi Akha E et al. Manuscript in preparation).

2.3.2. Marker choice in the HBB region

In the HBB region, I genotyped SNPs for which there were already assays designed and primers available in the Kwiatkowski lab. These SNPs were previously selected by Neil Hanchard (D.Phil thesis 2004) according to the following process: Project Ensembl (www.ensembl.org) was consulted to find SNPs spaced every 5 to 10 kb across 400 kb centred on the sickle mutation. There were 230 database SNPs submitted, of which 40 were validated. Nine SNPs in total had available frequency data, all of which were above 10%, but only one had frequency data in an African-related population (African Americans). As a result, SNPs were chosen on the basis of validation, preferably in an African population; available frequency data; and the desired SNP density. Two published SNPs (Currat,

Trabuchet et al. 2002) that were close to HbS but not in the database, were added to this set (HBB-707 and HBB-984). In all, 33 SNPs were chosen.

For my thesis, an extra five Restriction Fragment Length Polymorphism (RFLP) markers were chosen from literature to define the previously described classical β-globin haplotypes (Pagnier, Mears et al. 1984). Classical RFLP haplotypes were defined using 5 polymorphic restriction endonuclease sites within a 30-kilobase region of the β-globin-like gene cluster. For interpreting the haplotypic patterns of the 5 RFLP markers in terms of the classical haplotypes they describe in the β-globin like cluster (Table 2.3.2a), several publications were consulted (Pagnier, Mears et al. 1984; Steinberg, Lu et al. 1998; Rahimi, Karimi et al. 2003; Vivenes De Lugo, Rodriguez-Larralde et al. 2003).

	AvaII	HincII	Hind III in HBG1	Hind III in HBG2	Xmn I in HBG2
Benin	+	+	-	-	-
Car	+	-	-	+	-
Senegal	+	+	-	+	+
Arab	+	+	-	-	+
Cameroon	+	+	+	+	-

Table 2.3.2a: Classical HbS haplotypes designated by the 5 RFLP markers.
A + sign indicates the restriction Enzyme would cut, a – sign indicates absence of digestion by the enzyme at the site.

Primers were designed using the web-based Primer3 program© (available from the MIT-Whitehead website - http://www.broad.mit.edu/), and calculations of melting temperature (Tm) were checked using the group's web-based Tm calculator (K. Rockett). Primer sequences were blasted against the human genome sequence (BLAST - http://www.ncbi.nlm.nih.gov/BLAST/) to ensure selectivity. Primers were ordered from MWG Biotech® as HPSF purified at 0.01 μmol, reconstituted with sterile water to 100 μM, and labeled for storage on arrival.

Text cut off in original

Primers were designed to make sure they amplify a unique segment containing the targeted SNP, a process which involved careful consideration of the high degree of homology in the region due to genic duplication. It resulted in large fragments being amplified.

In total four fragments were required to be amplified to complete typing the 5 RFLP sites (Table 2.3.2b):

- HBG2 fragment (2734 bp in length) amplified the HBG2 gene and contained restriction sites for both Hind III and Xmn I RFLPs.
- HBG1 fragment (2909 bp in length) amplified the HBG1 gene and contained the restriction site Hind III.
- HBB fragment (1200 bp in length) amplified the HBB gene and contained the restriction site AvaII.
- The restriction site HincII was in an inter-genic region with unique flanking sequence, so a small fragment of 118 bp containing it was amplified.

RE	rs number	position	1st primer sequence	2ed primer sequence	fragment (in bp)
Ava II	rs10768683	chr11:5204367	AAATTAAGAAAAACAACAAC AAATGAATG	CATTCTAAACTGTACCCTGT TACTTATCC	1:
Hinc II	rs968857	chr11:5217034	ACGTTGGATGTCTGCCTCT GCTATAGTCTG	ACGTTGGATGCTGACTTCTG ATACTATGTCT	
Hind III	rs6578593	chr11:5226375	ACGTTGGATGCATGTACAC GCACATCTTATGTC	ACGTTGGATGCTTAAGAACC ATCCTTGCTACTCAG	2
Hind III	rs2070972	chr11:5231293	GACAGCATGAATACTTCCTG CC	ACGTTGGATGGAACTGAAGA CAACCATGTGTG	2
Xmn 1	rs7482144	chr11:5232745	ACGTTGGATGACAGCATGA ATACTTCCTGCC	ACGTTGGATGGAACTGAAGA CAACCATGTGTG	2

Table 2.3.2b: Positions of RFLP markers and primer sequences used to amplify PCR products in the HBB region, as well as the length of the amplified fragments.

2.4. Genotyping

Commonly used high-throughput SNP genotyping platforms include Sequenom (www.sequenom.com), Illumina GoldenGate (www.illumina.com), and Affymetrix ParAllele (www.affymetrix.com). The choice of which of these technologies to use depends on the focus of the study. The ultra-high throughput Affymetrix and Illumina platforms are more feasible to large genetic screens of many genes. They were the platforms used in genotyping MalariaGEN case-control samples.

The relatively limited set of polymorphisms genotyped in the 5q31 and HBB regions was undertaken using the Sequenom Matrix-Assisted Laser-Desorption/Ionosation Time-of-Flight Mass Spectrometry (MALDI-TOF MS) platform provided by the WTCHG Core Genomics group SNP typing service (www.well.ox.ac.uk/genomics). At the start of this project, only the homogenous MassExtend (hME) platform for low multiplexing (4- to 5-plex) was available.

2.4.1. hME platform

MassEXTEND is a SNP-typing procedure based on a primer-extension reaction for the polymorphism in question. It differentiates genotypes by allele-specific primer-extension. The MassEXTEND primer anneals adjacent to the polymorphic site and is extended dependent on the polymorphism. The hME platform uses a termination mixture with three dideoxynucleotides (ddNTPs), which terminate extension, and one deoxynucleotide (dNTP). For polymorphisms to be multiplexed together using a specific termination mixture, the extension products of the different alleles should differ by at least 15 Da in mass. An illustration of the allele-specific hME MassEXTEND process is presented in figure 2.4.1a. In this example, a ddATP/ ddCTP/ ddTTP/ dGTP termination mix resolves a C/T

polymorphism. The EXTEND primer (23-mer) anneals up to the polymorphic site and extends by one base on allele 1 (T) as a ddATP nucleotide is incorporated (24-mer), and by two bases on allele 2 (C) as a dGTP is incorporated, followed by a ddTTP nucleotide (25-mer). The 24-mer and 25-mer extended primers are differentiated using a mass spectrometer. Thus, where samples are homozygous for allele 1 (TT) or allele 2 (CC), a single peak is observed on the mass spectrometry trace at the predicted mass. Where a sample is heterozygous (CT), two peaks are observed on the mass spectrometry trace, one at the predicted mass for each allele. An additional "pause" peak might be observed on the trace. It represents the incorporation of one deoxynucleotide followed by incomplete extension by the polymerase. The pause peak is an artifact, but the position is taken into account for multiplex design, to avoid overlap with a true peak of another reaction.

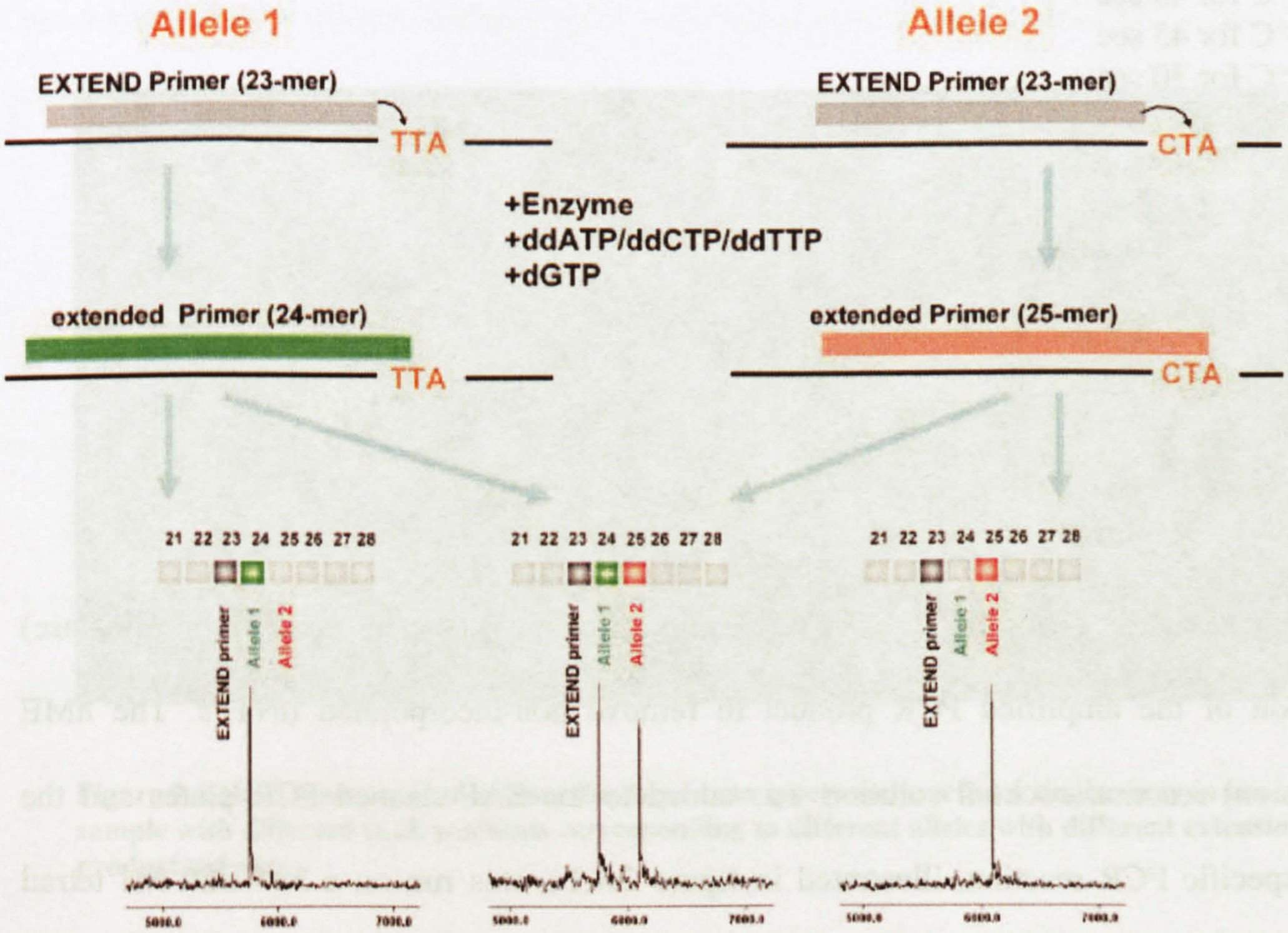


Figure 2.4.1a: Illustration of the massEXTEND process of the hME genotyping platform.
Image courtesy of Sequenom.

Amplification PCR

Sample DNA templates included PEP product diluted 1:20 with MilliQ water. A master mix consisting of reagents in the following volumes per sample was prepared; 0.25 μL dNTPs (8mM pooled); 0.2 μL $MgCl_2$ (50mM); 0.5 μL x10 PCR buffer; 0.025 μL Bioloine BioTaq (5U/μL); 1.025 μL MilliQ water; 0.01 μL forward primer (100μM); 0.01 μL reverse primer (100μM). Reaction mixture (3 μL) was dispensed into each well of an ABGene Thermofast 384-well plate. Of each diluted PEP sample, 2.0 μL were added to the reaction mixture in the corresponding well. PCR was undertaken on a 384-well MJ tetrad DNA engine under the following reaction conditions;

1. 96°C for 1:00 min
2. 96°C for 1:00 min
3. 94°C for 45 sec
4. 56°C for 45 sec
5. 72°C for 30 sec
6. go to 2 x5
7. 94°C for 45 sec
8. 65°C for 45 sec
9. 72°C for 30 sec
10. go to 6 x29
11. 72°C for 10:00 min
12. 15°C forever

Mass-EXTEND protocol

The Core Genomics group at the WTCHG undertook SAP (shrimp alkaline phosphatase) digestion of the amplified PCR product to remove non-incorporated dNTPs. The hME (extension) reaction cocktail solution was added to the SAP-cleaned PCR plates and the allele-specific PCR reaction, illustrated in figure 2.4.1a, was run on a 384-well MJ tetrad DNA engine. The hME product was treated with SpectroCLEAN resin to remove non-incorporated ddNTP's and dNTP's and salts. A SpectroPOINT robot spotted the cleaned product (15 nL) onto a 384 SpectroCHIP and calibrant was added onto calibrant patches of

the 384 SpectroCHIP. Genotypes were estimated by MALDI-TOF mass spectrometry. The quality of the runs was assessed by the Genomics team for the presence of probe peaks, clean spectra, unexpected peaks, and probe intensity. The resultant genotypes were presented in the form of 'Typer Analyzer' (figure 2.4.1b.), a software that displays the spectra for each run, highlights sample failures, and displays the predicted genotype calls for each sample and assay with an indication of the confidence of the call; conservative, moderate, aggressive, low probability, no alleles, or bad spectra.

Further details on Sequenom's hME genotyping process, including reaction cocktails and PCR cycle conditions can be found at the WTCHG Core Genomics website; http://www.well.ox.ac.uk/genomics/facilitites/sequenom/processing.shtml.

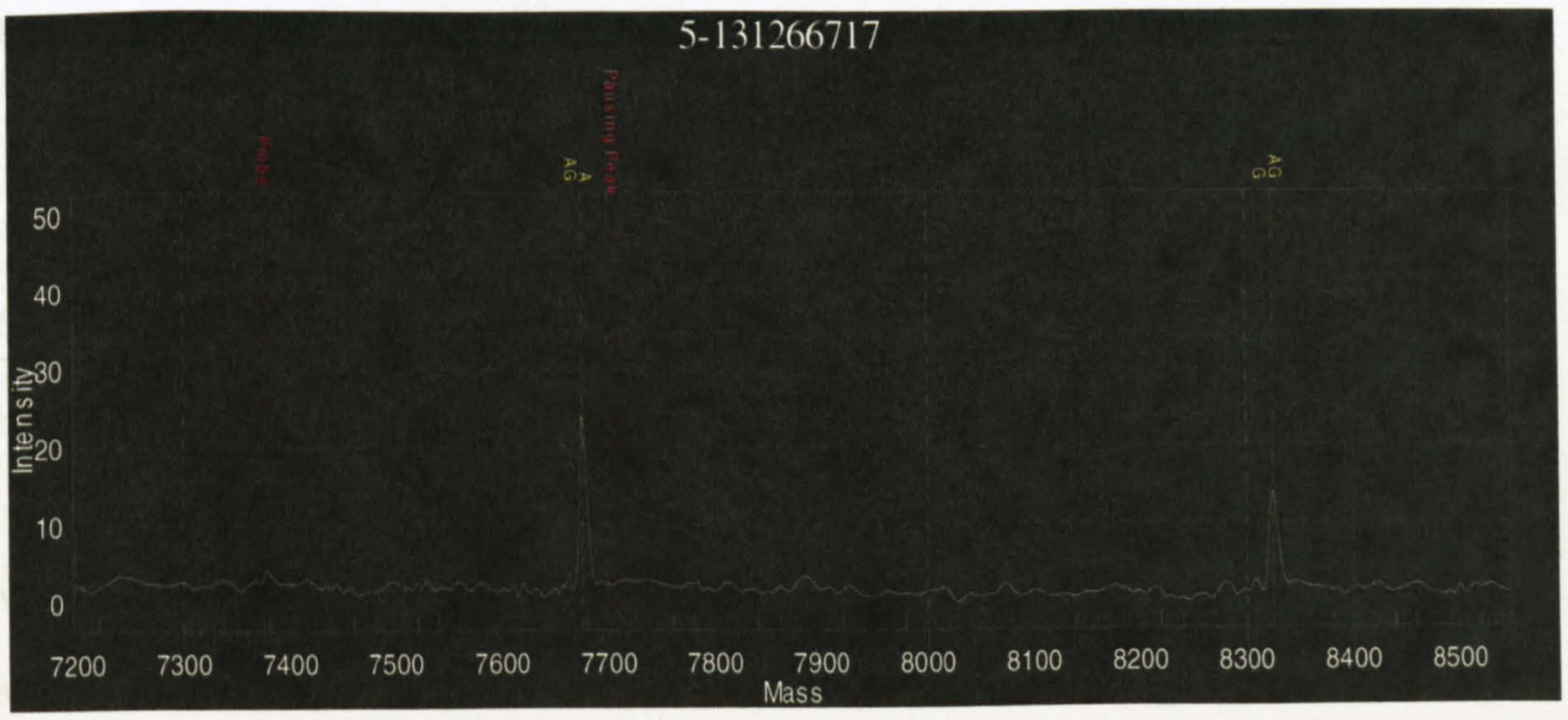


Figure 2.4.1b: A spectrum trace from the Sequenom typer analyzer for a single assay in one sample with different peak positions corresponding to different alleles with different extension product masses.

Sequenom hME primer design

Sequenom hME amplification and extension primers were designed using Sequenom's SpectroDESIGNER assay design software. Amplification primers were designed to produce an optimal amplicon size containing the polymorphic site at 80-120 bp. To avoid confusion in the mass spectrum, the mass of the amplification primer was designed to be different from the extension primer and its extension products. The SpectroDESIGNER software selected an appropriate termination mix for each SNP. The termination mix was selected such that DNA polymerisation terminated at the polymorphic site for one allele, (yielding an extended primer of one nucleotide), and as soon thereafter for the alternative allele. The SpectroDESIGNER software also grouped assays into optimal multiplexes. Criteria for assays within a hME multiplex included compatibility of extension primer peaks, minimal primer-primer interactions, and shared termination mix. For best results it was ensured that all extension primers and possible extension products differed by at least 15 Da, preferably 50 Da, with compromise for maximal multiplexing.

2.4.2. Genotyping the RFLP markers in the HBB region

2.4.2.1. PCR amplification

PCR reactions were carried out in a 96 well plates. Genomic DNA concentrations were standardised to 5 ng/ µL beforehand. A master mix consisting of reagents in the following volumes per 192 reactions was prepared; 44 µL $MgCl_2$ (50mM), 110 µL dNTPs (8mM pooled), 110 µL 10x PCR buffer, 5.5 µL Bioline Taq (5U/ µL), 386.1 µL MilliQ water, 2.2 µL forward primer (100µM); 2.2 µL reverse primer (100µM). The PCR mix was the same for all the fragments except the HBG1 fragment (2909 bp) for which a 3.3 µL of the forward and 3.3 µL of the reverse primers were used. Into each well of an ABGene Thermofast 96-

well plate, 6 μL of the reaction mixture were dispensed. Of each genomic DNA sample, 2.0 μL were added to the reaction mixture in the corresponding well. Each amplified fragment was done independently after optimizing the PCR conditions for it to get the best possible results for all the samples. PCR was undertaken on a 96-well MJ tetrad DNA engine under the following reaction conditions;

PCR protocol for HBG1 fragment:

1. 96°C for 1:00 min.
2. 94°C for 0:45 min.
3. 62°C for 2:30 min.
4. 72°C for 1:00 min.
5. go to 2 x5.
6. 94°C for 0:45 min.
7. 65°C for 2:30 min.
8. 72°C for 1:00 min.
9. go to 6 x29.
10. 72°C for 10:00min.
11. 15°C for 15:00 min.

PCR protocol for HBG2 fragment:

1. 96°C for 1:00 min.
2. 94°C for 0:45 min.
3. 64°C for 2:30 min.
4. 72°C for 1:00 min.
5. go to 2 x5.
6. 94°C for 0:45 min.
7. 65°C for 2:30 min.
8. 72°C for 1 min.
9. go to 6 x29.
10. 72°C for 10:00 min.
11. 15°C for 15:00 min.

PCR protocol for Hinc II fragment:

1. 96°C for 1:00 min.
2. 94°C for 0:45 min.
3. 56°C for 0:45 min.
4. 72°C for 0:30 min.
5. go to 2 x5.
6. 94°C for 0:45 min.
7. 65°C for 0:45 min.
8. 72°C for 0:30 min.
9. go to 6 x29.
10. 72°C for 10:00 min.
11. 15°C for 15:00 min.

PCR protocol for HBB fragment:

1. 96°C for 1:00 min.
2. 94°C for 0:45 min.
3. 56°C for 0:45 min.
4. 72°C for 1:00 min.
5. go to 2 x35.
6. 72°C for 10:00 min.
7. 15°C for 15:00 min.

2.4.2.2. Digestion by restriction endonuclease enzymes

Digestion enzymes and their buffers were ordered from New England BioLabs. Reaction mixes were prepared for each enzyme according to protocol as stated by the manufacturer.

Hind III digestion reaction:

Enzyme mix for one reaction

10x NEBuffer 2	1.5 µL
H_2O	9 µL
Hind III enzyme	0.5 µL

11 µL of enzyme mix added to 4 µL of PCR product.

Hinc II digestion reaction:

Enzyme mix for one reaction

10x NEBuffer	1 µL
H_2O	6.75 µL
Hinc II enzyme	0.25 µL

8 µL of enzyme mix added to 2 µL of PCR product.

XmnI digestion reaction:
Enzyme mix for one reaction

10x NEBuffer	1 μL
BSA	0.1 μL
H_2O	6.77 μL
XmnI enzyme	0.25 μL

8 μL of enzyme mix added to 2 μL of PCR product.

AvaII digestion reaction:
Enzyme mix for one reaction

1x NEBuffer 4	1 μL
H_2O	6.75 μL
AvaII enzyme	0.25 μL

8 μL of enzyme mix added to 2 μL of PCR product.

Digestion products loaded into an agarose gel and scored as (+ +) if the two alleles were digested, as (+ -) if one but not the other allele was digested (heterozygote), and as (- -) if no digestion occurred in the sample.

2.5. Statistical, analytical, and computational procedures

2.5.1. Analytical methods

2.5.1.1. Hardy-Weinberg Equilibrium

I used the Hardy-Weinberg equilibrium as an indicator of genotyping error rate. The underlying assumption is that where an assay experiences a low rate of genotype error, the observed genotype numbers should conform to the Hardy-Weinberg expected numbers. In the HWE model, the mathematical relation between the allele frequencies and the genotype frequencies is given by; AA: p^2 Aa: 2pq aa: q^2. Where, 'A' and 'a' are the major and minor alleles, respectively, and 'p' and 'q' are the allele frequencies of 'A' and 'a', respectively. The fit between the HWE model (expected) and the observed genotype

frequencies for each locus was tested with Pearson's chi-square test, with one degree of freedom.

2.5.1.2. Population Differentiation: Wright's Fixation Index (Fst)

Wright's Fst essentially provides a measure of the average reduction in heterozygosity within subpopulations relative to the total population. The Fst provides a measure of all effects of population substructure (at different levels of the population hierarchy) combined:

$$F_{ST} = \frac{Ht - Hs}{Ht}$$

Where Ht is the average HWE heterozygosity among organisms within the total area, and Hs is the average HWE heterozygosity among organisms within random-mating subpopulations. The theoretical minimum and maximum Fst values are 0 (indicating no genetic divergence) and 1 (indicating fixation for alternative alleles in different subpopulations), although Fst values of 1 are rarely observed. Wright proposed a rough guideline for the interpretation of Fst whereby values above 0.25 indicate very great genetic differentiation (Wright et al. 1978).

2.5.1.3. Linkage disequilibrium statistics

Estimates of pair-wise LD were, for the most part, derived from haplotypes. Two parameters were calculated; $|D'|$ and r^2 and are derived as follows. If a locus A has two alleles A and a, with major allele frequency f_A and minor allele frequency f_a, and a locus B has two alleles B and b, with major allele frequency f_B and f_b; then there are four possible haplotypes between

the two loci – AB, Ab, aB, ab whose frequencies can be denoted as f_{AB}, f_{Ab}, f_{aB}, and f_{ab} respectively. From this, the parameter D (Lewontin 1964) can be calculated as the difference between the frequency of any two-locus haplotype and the frequency those two alleles would be expected to show under random segregation (linkage equilibrium) such that,

$$D = f_{Ab} - f_A f_b$$

$$D = f_{ab} - f_a f_b$$

$$D = f_{AB} - f_A f_B$$

$$D = f_{aB} - f_a f_B$$

all of which are equivalent. The absolute value of D′ is calculated from D and D_{max}, the maximum theoretical value of D given the allele frequencies,

$$|D'| = \frac{D}{D_{max}}$$

where $D_{max} = \min(f_A f_B . f_a f_b)$ when D is negative and min $(f_a f_B . f_A f_b)$ when D is positive. This absolute value of D′, then, ranges from zero (linkage equilibrium – all haplotypes equally represented) to a maximum of one. The important characteristic of D′ is that if each of the four theoretical haplotypes is not observed in the sample, then the D′ statistic between those two markers will equal one.

The parameter r^2 is used for the correlation of the markers at the two sites and is given by

$$r^2 = \frac{D^2}{(f_A f_a f_B f_b)}$$

In practice r^2 is the same as the X^2 statistic (derived from a standard 2x2 table) divided by the number of chromosomes (Pritchard and Przeworski 2001). R-squared also ranges from a value of one (absolute and complete linkage – both alleles correlate perfectly and are at the same frequency) to zero. Although r^2 is less skewed by low allele frequencies and small sample sizes than D′, values equal to one are only very occasionally observed.

Calculation of LD was done using HaploXT (G. Abecasis – available from http://www.sph.umich.edu/csg/abecasis) and illustrated using **MARKER**, an in-house software programme that graphically plots pair wise LD measures (D. Kwiatkowski - available at http://www.gmap.net/marker).

2.5.2. Tools for detecting Signatures of positive selection

2.5.2.1. Haplotype-based tools for identifying signatures of recent positive natural selection

An alternative, and perhaps more powerful (Sabeti, Reich et al. 2002; Toomajian, Ajioka et al. 2003; Hanchard, Rockett et al. 2006), strategy to single locus tests of non-neutral evolution is haplotype-based analysis. Haplotype-based methods for detecting signatures of recent positive selection use information on the frequency of a given allele and the level of LD in the region surrounding that allele. Mutations under positive selection tend to increase in frequency at a sufficiently rapid rate that recombination does not substantially break down the haplotype on which the mutation arose. The reduction in the variability in sites surrounding selected mutations is known as a selective sweep. In contrast, alleles under neutral evolution tend to rise in frequency at a slow rate, over many generations, during which time recombination has substantially broken down the LD in the region surrounding the allele. Thus, a signature of positive selection is an allele with unusually long-range LD and high allele frequency. This has been demonstrated at numerous loci with previous

evidence of recent selection, including the *G6PD* (Sabeti, Reich et al. 2002), *LCT* (Johansson and Gyllensten 2008), *HBE* (Ohashi, Naka et al. 2004) and *HFE* (Toomajian, Ajioka et al. 2003) loci. I used two haplotype-based tests, the Long Range Haplotype (LRH) test implemented in *Sweep* (Sabeti, Reich et al. 2002), and the haplosimilarity score (Hanchard, Rockett et al. 2006) implemented in **MARKER**, to screen for putative signatures of recent positive selection in the 5q31 and HBB genomic regions, studied in the Sudanese samples.

2.5.2.2. Sweep

Sweep™ allows large-scale analysis of haplotype structure in genomes for the primary purpose of detecting evidence of natural selection. Primarily, it uses the LRH test to look for alleles of high frequency with long-range LD, which suggest the haplotype rapidly rose to high frequency before recombination could break down associations with nearby markers (Sabeti, Reich et al. 2002). *Sweep* takes phased genotype data as input, detects all haplotype blocks in that data, and then determines the frequency and long-range LD for each allele in each block.

http://www.broad.mit.edu/mpg/sweep/.

2.5.2.3. Haplosimilarity

The basis of the haplosimilarity test is essentially the same as that of the LRH test; high frequency alleles under neutral selection tend to be associated with a wider range of haplotypes than alleles under recent positive selection due to recombination. Thus, the two tests are highly correlated and have similar power (~80%) to detect reasonably strong selective sweeps even with a high recombination rate and minor allele frequencies below 0.2 (Hanchard, Rockett et al. 2006). However, while the LRH test searches for the interval at which haplotype homozygosity decays away from an allele of interest, the haplosimilarity test takes a predefined interval, or window, of SNPs and investigates the similarity of the

haplotypes within that window associated with an allele of interest. The haplosimilarity score provides a measure of haplotype similarity and is associated with the minor allele of the first SNP. All SNPs are investigated in turn by passing a sliding window across the region. A potential confounding effect due to variation in location, recombination and mutation rates is controlled for by comparison of the haplosimilarity score of a minor SNP allele relative to its major allele. haplosimilarity scores greater than 10 are considered relatively high (Hanchard, Rockett et al. 2006).

The Sudanese and HapMap SNP haplotypes for the 5q31 and HBB genomic regions were uploaded into **MARKER** (www.gmap.net), and the **MARKER** utilities were used to calculate the haplosimilarity scores.

2.5.3. Software for data analysis

2.5.3.1 Software for haplotype construction and interpretation

In diploid organisms, such as *Homo sapiens*, haplotypes are not distinct, and only unphased genotype data can be obtained through application of experimental techniques. Molecular haplotyping methods are available but these methods are not widely used because they incur significant costs and are low-throughput. Several algorithms for reconstructing haplotypes from unphased genotype data are now available (reviewed in (Niu 2004)). These algorithms offer practical, accurate, and cost-effective solutions. The difference in efficacy of haplotype reconstruction between algorithms tends to be modest when the region under study is short, as was generally the case in the candidate genes investigated here.

PHASE version 2.0

The **PHASE** programmes use a coalescence-based Markov-chain Monte Carlo (MCMC) approach: a pseudo-Gibbs sampler (PGS) for reconstructing haplotypes from genotype data

(Stephens, Smith et al. 2001). PGS uses Gibbs sampling to inform haplotype reconstruction from a priori expectations based on the coalescence theory. The approximate coalescent prior is based on the assumption that "the genetic sequence of a mutant offspring will differ only slightly from the progenitor sequence (often by a single-base change)" (Stephens and Donnelly 2003). **PHASE** version 2 is an advanced version of **PHASE**1, in which PGS algorithm has been slightly modified. Improvements in **PHASE**2 include allowance for recombination and decay in LD with distance.

Phamily-PHASE

PHASE (versions 1 and 2) requires genotypes from unrelated individuals as it uses population frequencies in the calculations. However, where additional family members have been genotyped, their genotypes can be used to infer the known haplotypes before running **PHASE**. This provides **PHASE** with more information enabling both more reliable results and faster execution. The phamily analysis is designed for one such situation. It takes a set of trio families and in the first stage uses logical methods only to infer all the known haplotypes in the parents. The children are then discarded and the parental genotypes and known haplotypes are passed to **PHASE** as a set of unrelated individuals. **PHASE** is used to estimate the most probable remaining haplotypes using statistical methods (Stephens, Smith et al. 2001).

MARKER beta

MARKER is a set of tools for exploring genetic markers in their genomic context (http://www.gmap.net/marker). The current beta version of **MARKER** only specifically deals with SNPs. **MARKER**'s key feature is a genome mapping tool that generates **MARKER** maps illustrating the LD between SNPs in a specified genetic region. A density of up to 200 SNPs (labelled by rs number) can be observed on a single map. Features of the

map such as the measure of LD, the SNP density, and the LD colour coding system can be altered as preferred. The main source of public data on **MARKER** is from the HapMap project (www.hapmap.org). However, **MARKER** maps can also be generated from private datasets. I used **MARKER** to generate LD maps of the 5q31 and HBB gene regions investigated here.

2.5.3.2. Software for detecting genetic differentiation between populations

The following software packages were used to detect and estimate the differences in the genetic makeup of different populations used in my studies.

ARLEQUIN

Arlequin software package (Schnieder et al. 2000) was used to assign individual genotypes to populations. This is done by determining the log-likelihood of each individual multi-locus genotype in each population, assuming that the individual comes from that population, taking into account the allele frequencies in each sample.

PHYLIP version 3.5c

Phylip software package (Felsenstein, J 1993) was used to construct the gene geneaologies: The haplotypes from different population samples were pooled together after phasing the genotypic data in each group separately. A distance matrix was calculated for the pooled haplotypes, and the algorithm UPGMA (Unweighted Pair-Group method with Arithmetic mean) was used to construct the phylogenetic gene tree that best describes the ancestral relationship between the haplotypes. When the population is divided in to two semi-isolated groups, for example, the alleles within each group are expected to be, on average, more similar to one another than comparisons between groups. This should produce two major clusters corresponding to the two populations in the constructed gene tree.

STRUCTURE version 2.1

The program ***STRUCTURE*** is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPs, microsatellites, RFLPs and AFLPs. The basic algorithm was described by Pritchard et al. (Pritchard, Stephens et al. 2000). Two models were used in the analysis: the no-admixture model, where the LD in the data is ignored. The other model is the linkage model, when any LD in the data is attributed to admixture in the population history. The models were provided with population-of-origin information for each individual.

KOIND

Using the KOIND package (Kosman and Leonard 2007), several within-population diversity measures are calculated (Nei(Hs), Muller(Mu), Kosman expected(K), Simpson(Si)). Values close to 0 indicate high uniformity, while large values indicate high diversity. A maximum of 200 bootstrap samples are generated by the software for each population's haplotypes. Measures of diversity are then averaged over all bootstrap-derived estimates. Several between-populations diversity measures can also be calculated using the same package (The Nei coefficient of differentiation(Gst), The Kosman-Leonard expected, The Rogers distance(R)). Values close to 0 indicate very little genetic differentiation. A maximum of 200 bootstrap samples are allowed to be generated for each population's haplotypes. Measures of genetic distance are then averaged over all bootstrap-derived estimates.

2.5.3.3. Software for bioinformatics and statistical analysis

All statistical analyses were carried out in either Microsoft Excel or R© version 2.7.0. http://www.r-project.org.

Perl scripts were compiled in UNIX and executed on Linux processors available in the Wellcome Trust Centre.

Chapter 3:

Genetic diversity, LD, and Haplotype structure of the 5q31 region in two ethnically distinct populations from neighboring villages in Eastern Sudan.

3.1. Abstract

The Hausa and Masalit of Eastern Sudan are two ethnically distinct but geographically contiguous populations that share the same environmental exposure to malaria and visceral leishmaniasis (VL). Epidemiological data suggest differential susceptibilities to these diseases in the two groups, with the Hausa appearing to be the more protected.

I chose to investigate whether there is any genetic basis for their differential susceptibilities by studying and comparing the genetic variation patterns, especially signals of recent positive selection, in one important genomic region – the 5q31- that has previously been implicated in both malaria and VL disease pathogenesis, and which hosts a large number of important immune genes that could be good candidates for malaria and/or leishmania susceptibility/resistance.

I genotyped 34 SNPs in 96 individuals from the Hausa village and 96 individuals from the Masalit village. Genetic diversity within and between groups was calculated, haplotypes phased, LD maps constructed, and signals of positive selection were tested for by available

metrics. I found these patterns not to be amenable to straightforward interpretations and concluded that they required further exploration.

3.2. Objectives

- Construct maps of LD and haplotype structure in the 5q31 region of the human genome, in the Hausa and Masalit of Eastern Sudan, and compare them to each other and to data from other populations.
- Quantify genetic diversity within each group. Discern and calculate the genetic distance between the two population groups.
- Look for signatures of positive selection in the 5q31 region and try to disentangle it from patterns created by the demographic history of these populations.

3.3. Introduction

3.3.1. Description of the Study area

The study area includes a set of villages that lie along the Rahad River, in Gadaref Province in eastern Sudan. The area has a total population of around 15,000, spread over 30-40 km along the river bank and consisting of 9 separate villages, each with its own distinct ethnic character. The major groups are Hausa, Masalit, Fulani, and Bergu. This study is carried out in two villages, **Koka** (population 1521), and **Salala** (population 1309).

3.3.2. Before association studies

Genetic association analysis is a popular approach for identifying genetic variations that correlates with phenotypic variation such as susceptibility to complex diseases, but there are numerous examples of associations that cannot be replicated or for which attempts to substantiate by linkage have failed, which has led to questioning the usefulness of the approach for common conditions (Terwilliger and Weiss 1998; Gambaro, Anglani et al. 2000; Weiss and Terwilliger 2000; Cardon and Bell 2001). It appears that understanding the structure of haplotypes in the human genome provides an important starting point for the study of complex traits. Haplotype methods have contributed to the identification of genes for Mendelian diseases, and recently, disorders that are both common and complex in inheritance (Hugot, Chamaillard et al. 2001; Rioux, Daly et al. 2001).

Because most association-study designs so far involve genotyping a small set of markers in genes or regions of interest and measuring their association with disease status they typically examine only a fraction of human genetic variation. Consequently, these studies rely on background marker correlations to detect disease association. With no knowledge of the properties of the variants studied (e.g. extent of genomic region for which variant provides information), it is difficult, to interpret their results (Cardon and Abecasis 2003).

In recent years, considerable effort has been directed towards elucidating the general properties of haplotypes in the human genome and underlying biological processes determining their variability. Some studies (Daly, Rioux et al. 2001; Patil, Berno et al. 2001) suggested a surprisingly simple pattern: blocks of variable length over which only a few common haplotypes are observed punctuated by sites at which recombination could be inferred in the history of the sample. One study showed that the boundaries of blocks and specific haplotypes they contain are highly correlated across populations and their results

suggested that haplotype blocks can be detected with a few markers (Gabriel, Schaffner et al. 2002) referred to as haplotype tagging SNPs (htSNPs). This notion of shared haplotype block boundaries and haplotypes across populations was the foundation and initial focus of the HapMap project (www.hapmap.org). But more recently, there has been a shift in the general consensus. It is now widely agreed that the generality of a block-like pattern is an oversimplification and is not a fundamental aspect of the genome (Liu, Sawyer et al. 2004; Sawyer, Mukherjee et al. 2005). Gu et al. (Gu, Pakstis et al. 2007) studied 10 loci in 38 diverse populations. They found considerable diversity in the pattern of LD, but nevertheless, very high transferability of tagSNPs was also found.

Another major area of investigation, implicated in designing and interpreting association studies is population structure. Most studies of human variation begin by sampling from predefined "populations." These populations are usually defined on the basis of culture or geography and might not reflect underlying genetic relationships. Self-reported ancestry can facilitate assessments of epidemiological risks but does not obviate the need to use genetic information in genetic association studies. Uncorrected population stratification may lead to false positives in association studies when there are systematic differences in the ancestry of cases and controls.

The problems associated with very tight LD in regions that have been identified as being both linked and associated with disease, is that many alleles in a gene or genes might be strongly associated, thereby precluding resolution of individual effects. Given strong evidence for LD, it can prove difficult or impossible to identify causative mutations in disease genes themselves. This can be further complicated by the clustering of loci that share similar functions within a single genomic region of strong LD. One approach to breaking down such regions of linkage disequilibrium is to characterize the disease phenotype in diverse populations that might share substantially different ancestries for the genomic region

of interest. Such 'trans-racial mapping' has allowed the dissection of strongly conserved regions with extensive LD (Cardon and Bell 2001).

3.3.3. Evidence of differential malaria susceptibility across populations

The structure of human populations is relevant in various epidemiological contexts. As a result of variation in frequencies of both genetic and nongenetic risk factors, rates of disease and of such phenotypes as adverse drug response vary across populations. Differences in susceptibility to malaria between ethnic groups have previously been observed in many studies (Bryceson, Fleming et al. 1976) (Greenwood, Groenendaal et al. 1987; Terrenato, Shrestha et al. 1988). From studies on sympatric ethnic groups in Burkina Faso, the Fulani were found to be more resistant to malaria, exhibiting fewer clinical attacks and lower parasitaemia, than the Mossi and Rimaibe (Modiano, Petrarca et al. 1996). The enhanced resistance of the Fulani appears to reflect genetic factors. It has been demonstrated that the Fulani have high levels of antimalarial antibodies (Modiano, Chiucchiuini et al. 1998) and a low frequency of protective globin variants and other classical malaria resistance genes (Modiano, Luoni et al. 2001).

The apparently less severe malaria observed in the Hausa compared to Masalit could well be due to a protective mechanism similar to that reported in the Fulani of West Africa, particularly when considering the shared genetic, cultural and political history between Hausa and Fulani.

3.3.4. Why we are interested in the 5q31-33 region

In a candidate-region approach to the human genetics of *P. falciparum* infection levels, chromosomal regions that contain genes involved in immune responses are of major interest. Chromosome 5q31-q33 region contains numerous candidate genes encoding immunological molecules such as cytokines, growth factors, and growth-factor receptors which are involved in the control of immunity to *P. falciparum* blood stages, in particular a cluster of candidate genes coding for the CSF2, IL-3, IL-4, IL-5, IL-13, and IRF-1 that regulates IFNγ transcription. A large number of observations indicated that IFNγ is critical in immunity against intracellular pathogens. Among candidates that map in the distal part of 5q31 is IL12p40 which encodes the β chain of IL12 which has been shown to protect monkeys against P.cynomolgi sporozoite induced infection (Garcia, Marquet et al. 1998; Rihet, Traore et al. 1998). More recently, the involvement of the IRF-1 locus within the 5q31 region with malaria susceptibility has been established in a case control study in the Mossi ethnic group from West Africa (Mangano, Luoni et al. 2008).

Moreover, the importance of this region in immune regulation is highlighted by its linkage to plasma immunoglobulin E (IgE) levels (Marsh, Neely et al. 1994; Meyers, Postma et al. 1994), bronchial hyperresponsiveness (Postma, Bleecker et al. 1995), and schistosomiasis infection (Marquet, Abel et al. 1996).

3.3.5. The genomic region approach

Identification of patterns of LD at the genomic level as well as within specific genes is useful for mapping genes associated with complex diseases. Knowledge of variant frequencies and their relationships can reduce the uncertainty in the design and interpretation of association studies. It can act as a guide to predict adequate coverage of a particular

genomic region and help point to the probable location of functional variants when association to a SNP marker is identified.

This approach can yet be refined by the detection of background signals of selection in the genomic area of interest (Hamblin and Di Rienzo 2000; Saunders, Hammer et al. 2002; Bamshad and Wooding 2003). It is important to model selection in order to understand patterns of allele frequency variation, including haplotype frequencies in data from studies in genetic epidemiology. In theory it may be possible to identify putative genetic disease factors by identifying regions of the human genome that are currently under selection.

As the signals of positive selection are often confounded by demographic factors, an understanding of population history is crucial for identifying the genes that are subject to selection. And because of the effect of demographic assumptions on the population genetic neutrality tests, it would not be very meaningful to reject the standard neutral model using these methods without paying careful attention to the underlying demographics.

In conclusion, knowledge of the patterns of genetic diversity, LD and haplotype structure, as well as patterns of positive selection in the 5q31 region across sympatric groups, and populations in other parts of the world, has important implications for the identifications of SNPs and haplotypes useful for genetic mapping studies of susceptibility to complex diseases.

3.4. Materials and Methods

3.4.1. Sampled populations and study area

Sampling was carried out in two villages along the eastern bank of Rahad river area of eastern Sudan along the Sudanese-Ethiopian border, 400 km south-east of Khartoum. Koka

village is 35 km north of Salala village. This area is the major endemic area for visceral leishmaniasis in Sudan (75% of reported cases in 1987), it is also endemic for malaria mainly *P.falciparum* infection.

3.4.2. DNA collection and preparation

The study was reviewed and ethically approved by the Ethical Committee of the Institute of Endemic Diseases, University of Khartoum. Samples were taken with informed consent from all individuals (see Appendix 1).

With the help of other members of the Institute of Endemic Diseases, I collected DNA samples using the buccal brush method, and extracted DNA by the guanidine hydrochloride method (see Materials and Methods chapter). Total yield for a sample was 20 μg on average. At the WTCHG in Oxford, I quantified and archived the DNA samples. First, sample DNA concentration was quantified using the PicoGreen assay (Molecular Probes, Leiden, Netherlands). DNA samples were then prepared to give 20ng/μL stock samples. Whole genome amplification was carried out using Primer Extension Pre-amplification PCR Method (PEP) (Zhang, Cui et al. 1992).

3.4.3. Choice of markers

Thirty four markers were chosen from a larger set of markers in the 5q31 region that had previously been tested in the laboratory at the WTCHG in Oxford, in samples from The Gambia and the UK (see Materials and Methods chapter for details). These 34 were selected as the most efficient set of markers to capture most of the haplotypic diversity in those populations (haplotype tagging SNPs). These SNPs are listed in table 3.4.3. Their distribution and relation to genes in the region is displayed in figure 3.4.3.

ID number	Assay (rs number)	Chromosomal coordinate	Gene	Alleles
1	rs12656759	131395509	ACSL6	A/G
2	rs3091334	131419372	IL3	C/T
3	rs31473	131432345	CSF2	A/T
4	rs27348	131435038	CSF2	A/T
5	rs1469149	131436741	CSF2	A/C
6	rs27438	131441154	CSF2	A/G
7	rs162881	131636693	PDLIM4	G/T
8	rs157572	131654011	ENSG00000205179	C/G
9	rs272842	131684416	SLC22A4	A/G
10	rs274559	131747969	SLC22A5	C/T
11	rs274549	131757017	SLC22A5	G/T
12	rs12517950	131759225	SLC22A5	C/T
13	rs17689595	131760673	SLC22A5	C/T
14	rs11739135	131761296	SLC22A5	C/G
15	rs1016988	131772473	NP_001013739.1	A/G
16	rs12521868	131812292	NP_001013739.1	A/C
17	rs2522044	131819745	NP_001013739.1	A/G
18	rs2522057	131829846	NP_001013739.1	C/G
19	rs7719499	131841990	NP_001013739.1	C/G
20	rs2070727	131848174	IRF1	A/C
21	rs10068129	131849145	IRF1	C/T
22	rs2706384	131854779	IRF1	A/C
23	rs2069820	131904015	IL5	C/G
24	rs17166050	131943112	RAD50	C/T
25	rs12187537	131967803	RAD50	A/C
26	rs3798135	131993008	RAD50	A/G
27	rs1800925	132020708	IL13	A/G
28	rs1295686	132023742	IL13	A/G
29	rs20541	132023863	IL13	C/T
30	rs1295685	132024344	IL13	C/T
31	rs2243219	132030024	IL13	A/G
32	rs734244	132038625	IL4	A/G
33	rs2227284	132040624	IL4	A/C
34	rs2243270	132042008	IL4	C/T

Table 3.4.3: SNPs typed in the 5q31 in the Sudanese samples. Their coordinates in chromosome 5 are shown (Ensemble version 39). Also shown are the closest genes to the polymorphisms and their alleles.

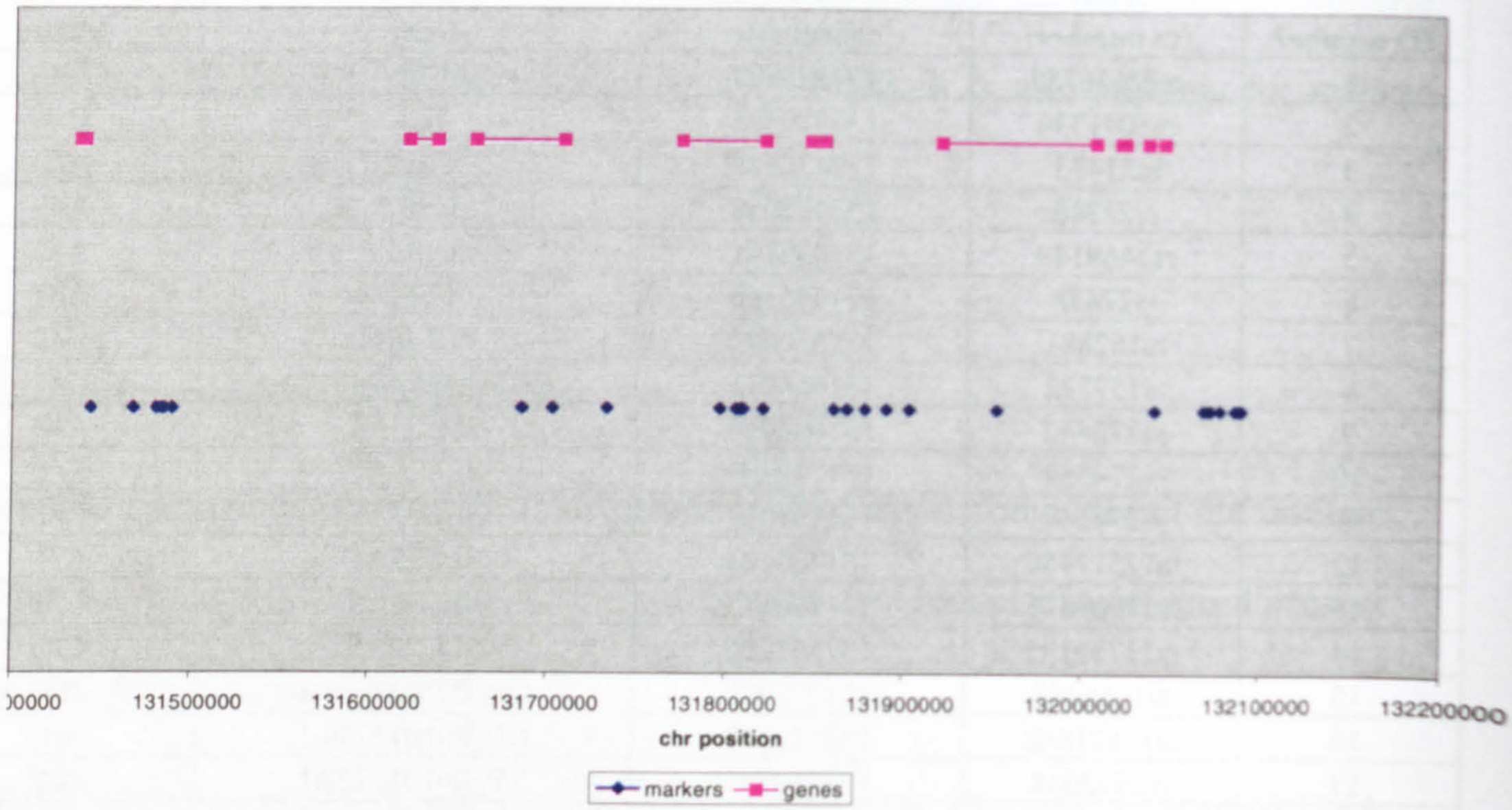


Figure 3.4.3: The distribution of SNPs that were typed in the 5q31 in the Sudanese samples and their relation to genes in the region. The SNPs and genes are ordered from left to right as in table 3.4.3.

3.4.4. Genotyping the 5q31 genomic region

The genomic area typed spanned a 646.5 kb in the 5q31 region (from coordinate 5:131395509 to coordinate 5:132042008 according to Ensemble version 39). From each village 96 individuals were genotyped. In the Masalit, 63 individual comprised trios of mother, father, and child. 12 individuals were in parent-child pairs, and the 21 remaining individuals were unrelated. In the Hausa; 42 individual comprised trios of mother, father, and child. 32 individuals were in parent-child pairs, and the 19 remaining individuals were unrelated. Genotyping was carried out using primer-extension/mass-spectrometry (Sequenom, San Diego, CA, USA) technology (see Materials and Methods chapter). There was a very high genotyping success rate of 97%.

3.4.5. Statistical, analytical, and computational procedures

HWE calculations for assay genotyping performance were undertaken in excel. A compromise between the exclusion of interesting variants with significant genotype distortion from HWE due to the selective pressure of malaria and the exclusion of inaccurate genotyping assays, a 0.1% HWE chi-square significance threshold was set for assay exclusion.

3.4.5.1. Haplotype construction

Haplotypes were generated from the genotypic data of unrelated individuals using the **Phamily-PHASE** and **PHASE** ***version2.1*** software packages (Stephens, Smith et al. 2001; Stephens and Donnelly 2003). Because I typed additional family members, their genotypes were used to infer the known haplotypes before running **PHASE**. This provided **PHASE** with more information enabling both more reliable results and faster execution. The Phamily analysis is designed for one such situation. It takes a set of trio families and in the first stage uses logical methods only to infer all the known haplotypes in the parents. The children were then discarded and the parental genotypes and known haplotypes were passed to **PHASE** as a set of unrelated individuals. **PHASE** was used to estimate the most probable remaining haplotypes using statistical methods.

3.4.5.2. LD maps and signals of positive selection

Haplotypes of unrelated individuals generated from **PHASE**, were fed – each population group separately- into the ***MARKER*** application (http://www.gmap.net/marker). **MARKER** maps were generated illustrating the LD between SNPs in the 5q31 genetic region. I chose the disequilibrium coefficient D' as the LD measure for these maps. Haplosimilarity (Hanchard, Rockett et al. 2006) was computed by the application as a way for detecting putative signatures of recent positive selection.

3.4.5.3. Detecting genetic differentiation between the Hausa and Masalit

The following software was used to detect and estimate the differences in the genetic makeup of the Hausa and Masalit (see Material and Methods chapter for details)

ARLEQUIN software package (Schnieder et al. 2000) was used to assign individual genotypes to populations. This is done by determining the log-likelihood of each individual multi-locus genotype in each population, assuming that the individual comes from that population, taking into account the allele frequencies in each sample.

PHYLIP version 3.5c software package (Felsenstein, J 1993) was used to construct the gene geneaologies: The haplotypes from the Hausa and Masalit samples were pooled together after phasing the genotypic data in each group separately.

***STRUCTURE v2.1*:** (Pritchard, Stephens et al. 2000) Analysis was carried out with 100,000 burn-in and 100,000 iterations. Two models were used in the analysis: the no-admixture model, where the LD in the data is ignored, assuming two populations of origin. The model was provided with population-of-origin information for each individual. The other model is the linkage model, when any LD in the data is attributed to admixture in the population history. The linkage model was run using the phased haplotypes of the unrelated individuals and providing population-of-origin information. For estimating K, 1,000,000 iterations were used for assumed number of populations (k) between 1 and 10.

3.5. Results

3.5.1. Checking for pedigree errors

Twenty-nine of the 34 typed markers were typed successfully and found to be polymorphic in the two populations. For 11% of the sampled individuals, their pedigree did not concur with their genotypic data. For every trio with two or more markers with pedigree inconsistencies out of a possible 29, the three individuals making the trio were analysed as unrelated. After the pedigree check there were 72 unrelated Masalit, and 72 unrelated Hausa individuals.

3.5.2. Assay Properties

Table 3.5.2a and table 3.5.2b present features of the 5q31 assays genotyped in the Hausa and Masalit samples respectively, including their minor allele frequencies in the unrelated individuals –after excluding the children-, their genotyping performance in terms of failure rate (%), and conformation of observed genotype distributions to the expected HWE distribution. The assays generally performed well on the MALDI-TOF mass spectrometry platform. All assays exhibited low failure rates, the majority falling below 5%, and none exceeded 15%. Typed assays demonstrated high concordance with HWE. They were all within the 0.1% significance threshold. This threshold, rather than the typical 5% threshold, was defined in the attempt to discriminate between poor genotype call rate and effect of selection pressure.

Assay (rs number)	Minor allele frequency	% Failure	*HWE chi-square
rs12656759	0.49	0.04	0.7(0.40)
rs3091334	0.32	0.11	0.68(0.41)
rs31473	0.49	0.00	1.39(0.24)
rs27348	0.23	0.01	0.01(0.91)
rs1469149	0.42	0.01	0.02(0.90)
rs27438	0.39	0.01	2.29(0.13)
rs162881	0.46	0.04	0.01(0.94)
rs157572	0.46	0.01	0.41(0.52)
rs272842	0.34	0.03	0.01(0.90)
rs274559	0.35	0.03	0.09(0.76)
rs274549	0.42	0.00	0.27(0.60)
rs12517950	0.04	0.01	0.09(0.76)
rs11739135	0.04	0.01	0.09(0.76)
rs1016988	0.09	0.04	0.75(0.39)
rs12521868	0.05	0.00	0.19(0.66)
rs2522044	0.22	0.04	3.88(0.05)
rs2522057	0.06	0.06	0.27(0.61)
rs7719499	0.48	0.00	10.45(0.0012)
rs2706384	0.42	0.01	6.84(0.01)
rs2069820	0.02	0.04	0.03(0.85)
rs3798135	0.37	0.01	0.57(0.45)
rs1800925	0.38	0.01	0.02(0.89)
rs1295686	0.30	0.00	0.79(0.37)
rs20541	0.13	0.10	0.94(0.33)
rs1295685	0.04	0.06	0.1(0.75)
rs2243219	0.41	0.00	0(0.97)
rs734244	0.48	0.15	0.14(0.71)
rs2227284	0.06	0.03	0.33(0.57)
rs2243270	0.27	0.00	0.58(0.44)

Table 3.5.2a: Genotyping Performance of *5q31* SNPs in 72 Unrelated Hausa Individuals (144 Chromosomes). * Hardy-Weinberg equilibrium chi-square value with probability (p-value) in brackets.

Assay (rs number)	Minor allele frequency	% Failure	*HWE chi-square
rs12656759	0.29	0.03	2.53(0.11)
rs3091334	0.27	0.07	3.87(0.05)
rs31473	0.44	0.03	1.24(0.27)
rs27348	0.21	0.00	2.31(0.13)
rs1469149	0.41	0.03	0.24(0.63)
rs27438	0.42	0.00	1.01(0.32)
rs162881	0.40	0.06	0.42(0.52)
rs157572	0.49	0.00	0.00(0.99)
rs272842	0.14	0.00	1.87(0.17)
rs274559	0.14	0.00	1.87(0.17)
rs274549	0.35	0.01	0.01(0.92)
rs12517950	0.02	0.00	0.03(0.86)
rs11739135	0.02	0.01	0.03(0.86)
rs1016988	0.15	0.00	2.1(0.15)
rs12521868	0.03	0.00	0.06(0.81)
rs2522044	0.20	0.01	2.53(0.11)
rs2522057	0.04	0.01	0.14(0.71)
rs7719499	0.39	0.00	9.11(0.003)
rs2706384	0.33	0.08	6.26(0.01)
rs2069820	0.02	0.01	0.03(0.86)
rs3798135	0.42	0.00	0.06(0.81)
rs1800925	0.38	0.00	1.55(0.21)
rs1295686	0.17	0.01	2.78(0.1)
rs20541	0.20	0.01	0.00(0.98)
rs1295685	0.01	0.03	0.00(0.95)
rs2243219	0.46	0.00	0.29(0.59)
rs734244	0.48	0.04	2.41(0.12)
rs2227284	0.01	0.03	0.01(0.9)
rs2243270	0.14	0.00	1.87(0.17)

Table 3.5.2b: Genotyping Performance of *5q31* SNPs in 72 Unrelated Masalit Individuals (144 Chromosomes). * Hardy-Weinberg equilibrium chi-square value with probability (p-value) in brackets.

3.5.3. Comparing allele frequencies in the Hausa and Masalit

To compare allele frequencies between the Hausa and Masalit in the 5q31 region, I calculated the allele frequencies in each population group of unrelated individuals. Each marker was then compared between the two groups using a 2x2 chi square test with one degree of freedom. Although the differences of allele frequencies between the groups were mostly insignificant (Figure 3.5.3a), the sampled Hausa individuals tended to be higher in

their minor allele frequency compared to the Masalit sample. One possible explanation for this minor difference is that the Hausa might have a greater effective population size due to higher migration and integration of individuals from nearby Hausa villages through marriage. If that were true, they could be less influenced by genetic drift which, generally, tends to sweep gene variants out of a population over time.

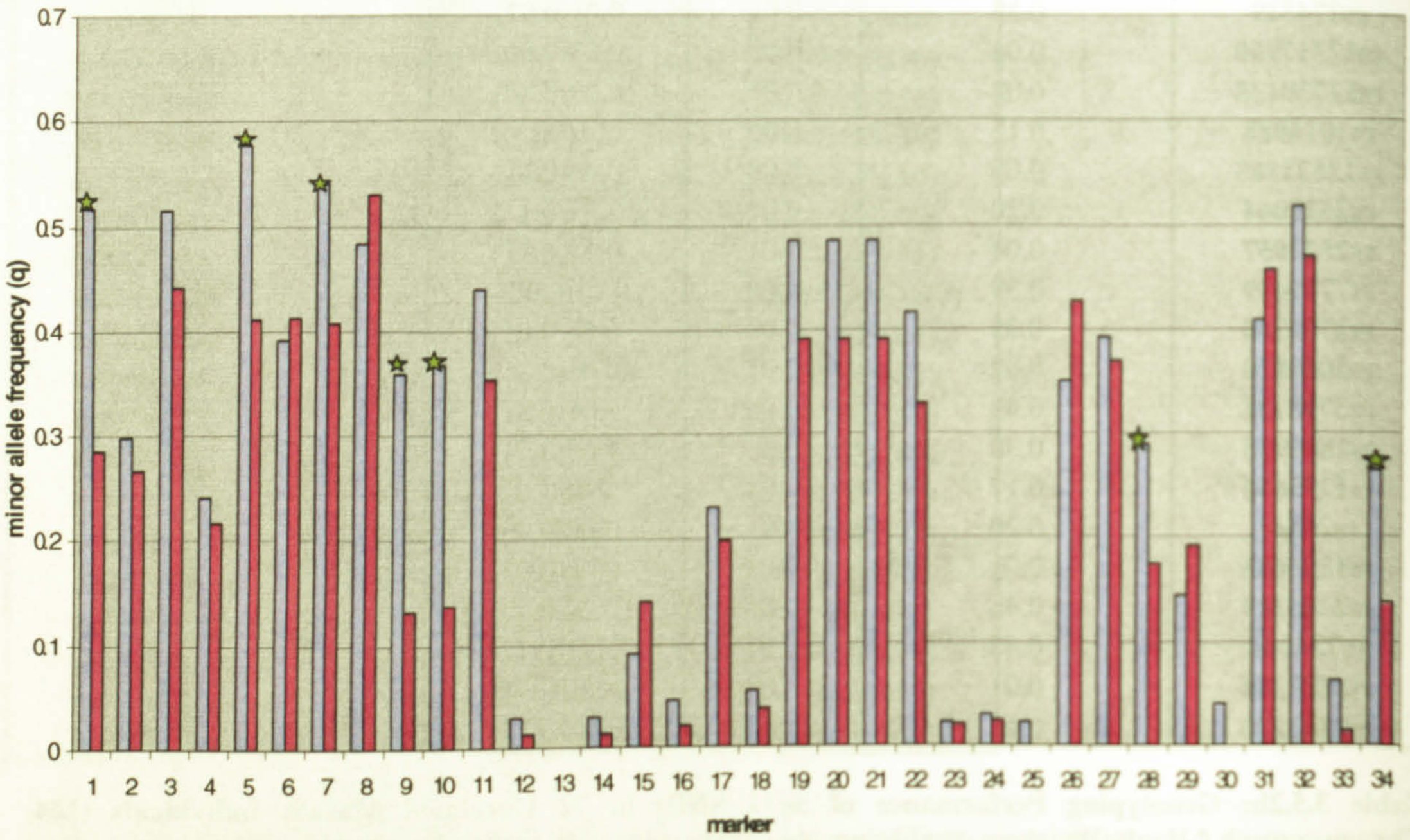


Figure 3.5.3a: Minor allele frequencies of 5q31 markers typed in the Hausa (blue bars) and Masalit (purple bars). Stars indicate the markers that were found to be significantly different between the two populations. Markers are ordered on x axis as shown in Table 3.4.3. Minor allele frequencies (q) are shown on the y axis.

Minor allele frequencies were found to be highly correlated between the Hausa and Masalit samples, with correlation coefficient $R^2 = 0.81$ (figure 3.3.3b).

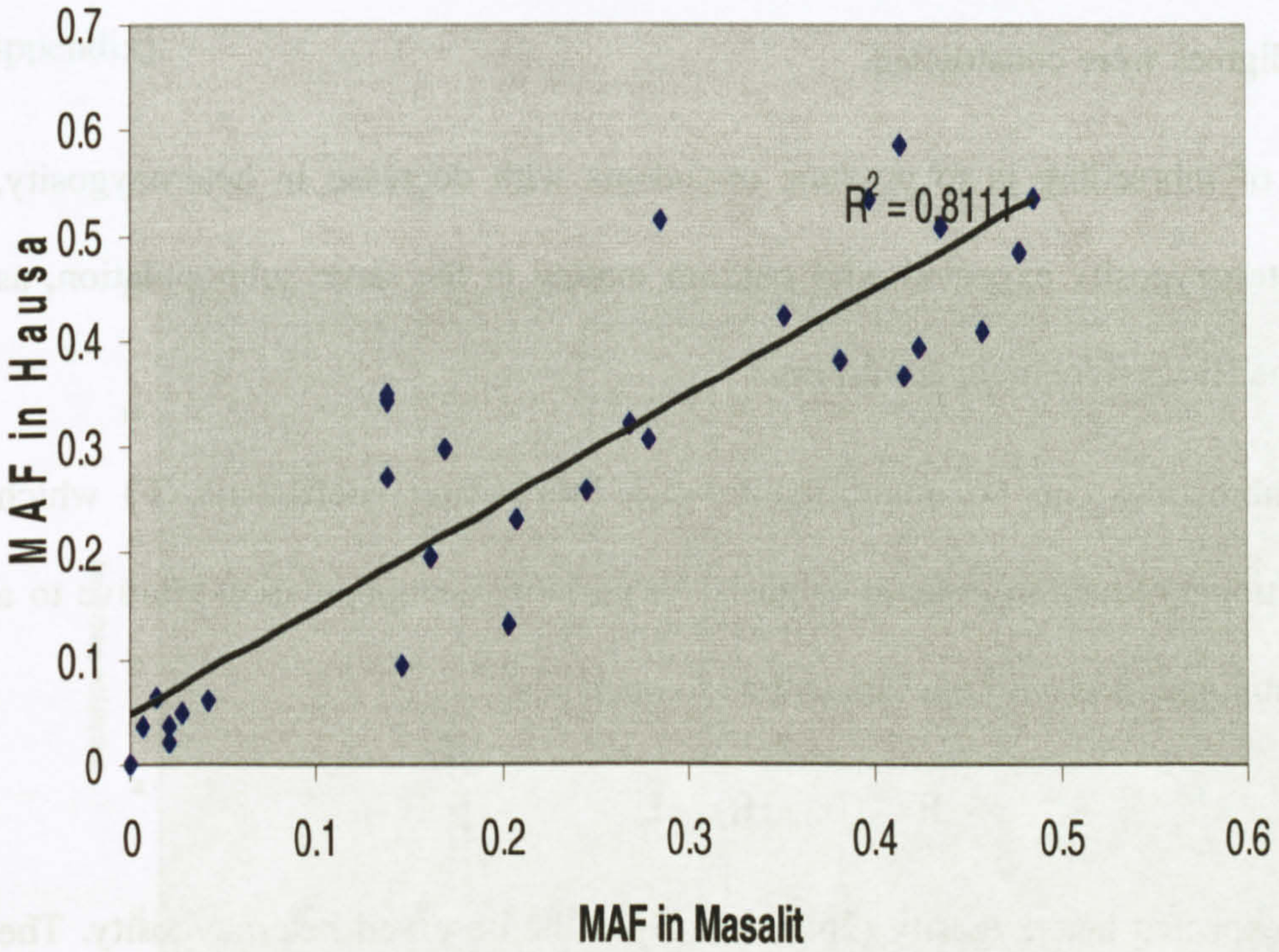


Figure 3.5.3b: Correlation of minor allele frequencies (MAF) in the Hausa and Masalit for markers typed in the 5q31 region.

To determine whether this level of correlation in MAF values is greater than might be expected for any two different ethnic groups in Africa, more population groups should be analysed. The question of whether this observation is specific to the 5q31 region or not, could be tackled by looking at other genomic regions in the Hausa and Masalit.

3.5.4. Inbreeding

When mating takes place between relatives, the pattern of mating is called inbreeding. In the two Sudanese populations, I expected to find some evidence of inbreeding, because of the fact that most individuals were found to be related to some degree within each village when whole-village pedigrees were constructed.

The main effect of inbreeding is to produce organisms with decrease in heterozygosity, relative to the heterozygosity expected with random mating in the same subpopulation, as calculated from the Hardy Weinberg equilibrium.

The effects of inbreeding can be quantified by **the inbreeding coefficient (F)** which measures the fractional reduction in heterozygosity of an inbred subpopulation relative to a random mating subpopulation with the same allele frequencies.

$$F = (H_0 - H_1) / H_0$$

Where H_0 is the expected heterzygosity (2pq), and H_1 is the observed heterozygosity. The value of F equals zero when there is no inbreeding. For the Masalit sample the average F over all typed markers was (-0.04235) and for the Hausa (-0.04568). This suggests either an inadequacy or a bias in the samples' choice. The rigorous criteria used for choosing unrelated individuals might have resulted in samples that don't represent the villages' pedigree structure, with a disproportionate representation of individuals from outside the villages.

3.5.5. Haplotype Analysis

Using the software package **PHASE 2.1** (Stephens, Smith et al. 2001; Stephens and Donnelly 2003) to infer the chromosomal phase of the parental genotypes. Haplotypes were

generated by integrating family- and population-based reconstruction methods. The Masalit were found to have 117 distinct haplotypes across the region, out of a possible 144. Only two haplotypes were common (frequency > 5%), the highest had a frequency of 11 identical copies (figure 3.5.5a) (For a full list of haplotypes sequences and **PHASE** probabilities see appendix).

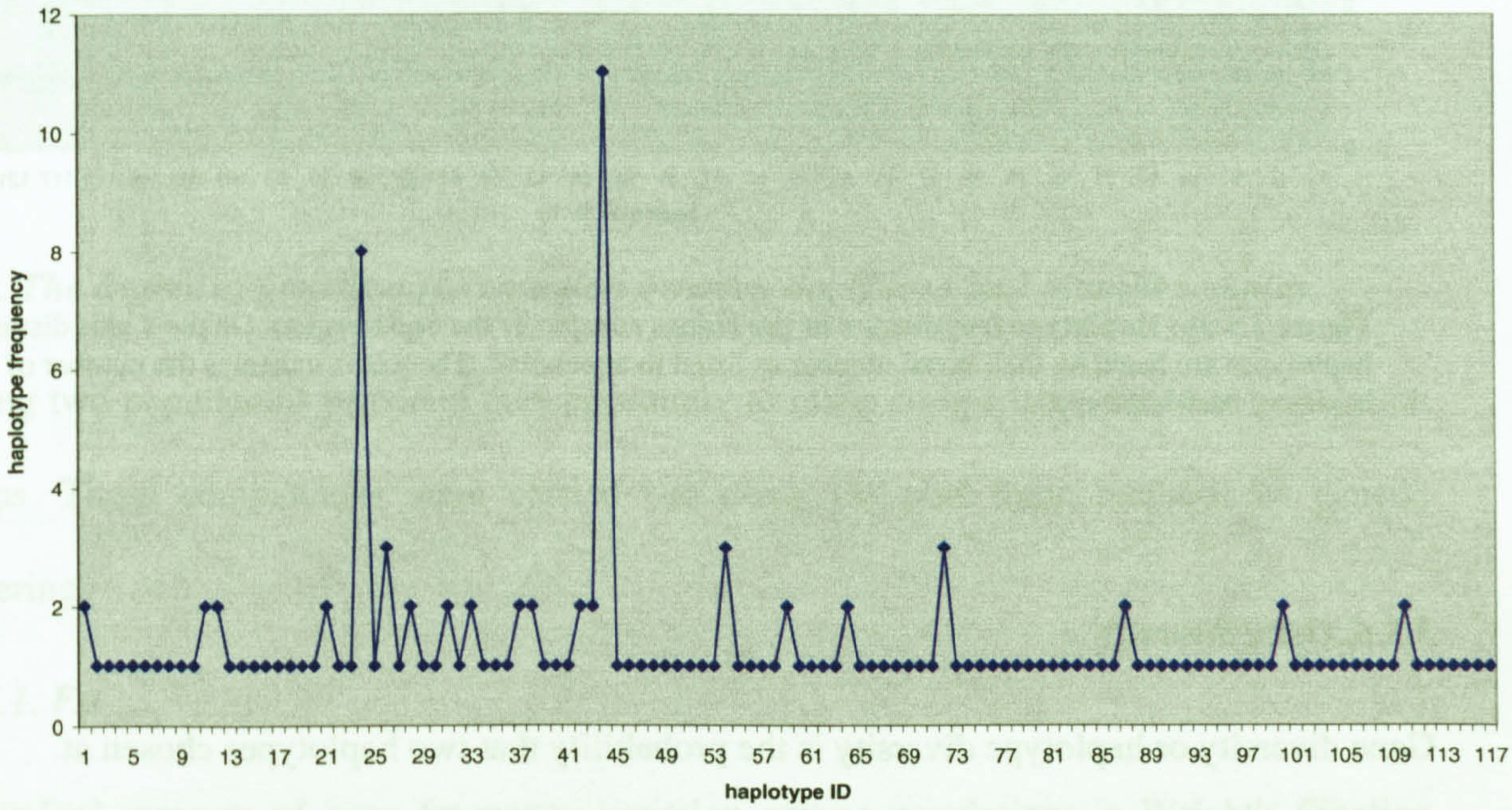


Figure 3.5.5a: Haplotype frequencies in the Masalit sample in the 5q31 region. On the x axis distinct haplotypes are listed by their serial number as listed in appendix2. The y axis indicates the number of copies of each haplotype.

The Hausa were found to have 127 distinct haplotypes across the region, out of a possible 144. Only two haplotypes were common (frequency > 5%) in this sample as well, the highest had a frequency of 9 identical copies (figure 3.5.5b) (For a full list of haplotypes sequences and **PHASE** probabilities see appendix).

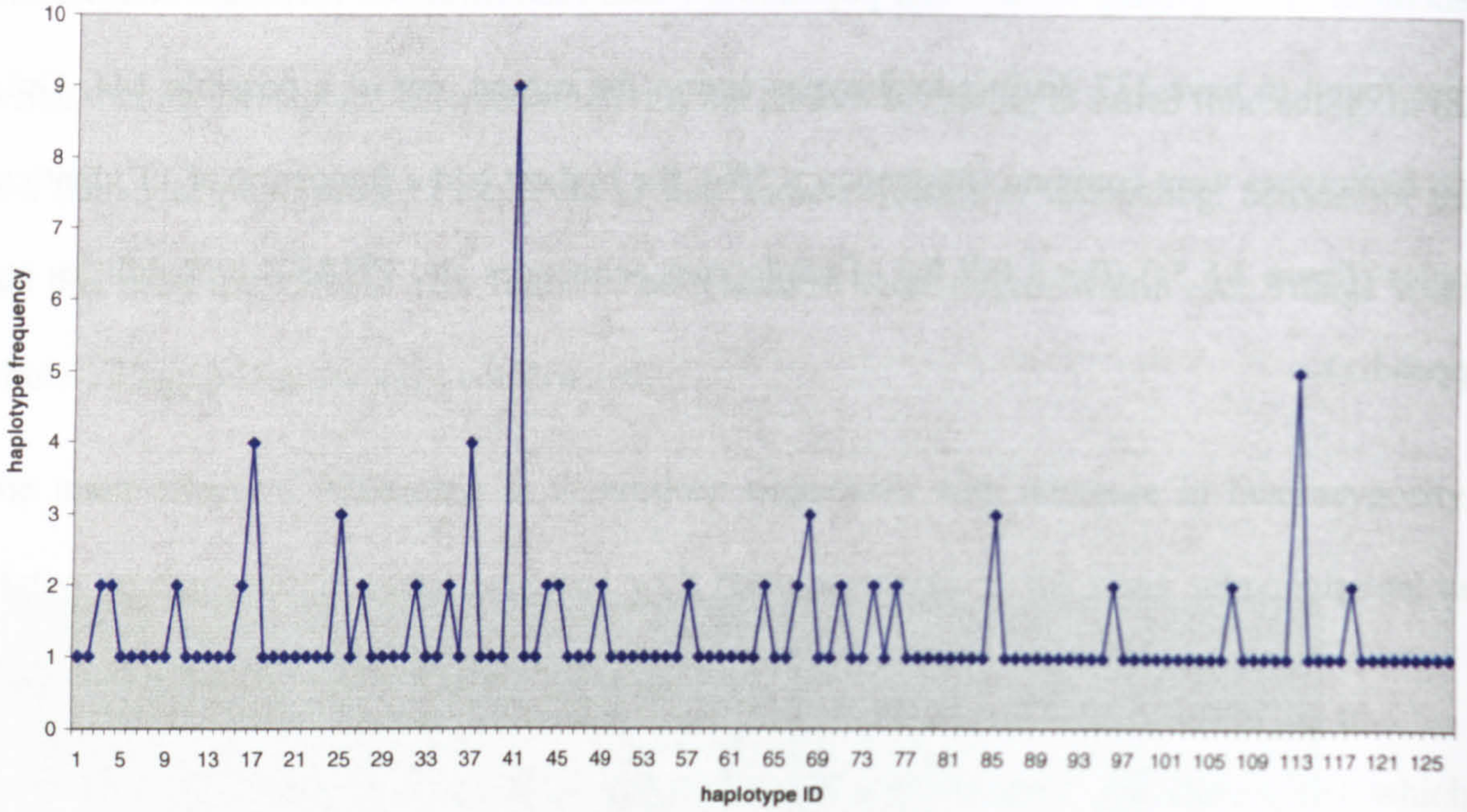


Figure 3.5.5b: Haplotype frequencies in the Hausa sample in the 5q31 region. On the x axis distinct haplotypes are listed by their serial number as listed in appendix2. The y axis indicates the number of copies of each haplotype.

3.5.6. Gene diversity

Gene diversity or haplotype diversity is the probability that two haplotypes chosen at random from the sample are different.Haplotype diversity was calculated using Nei formula (Nei 1978) as follows:

$$H = \frac{n}{n-1}\left(1-\sum_{i=1}^{k} p_i^2\right)$$

Where n is the number of chromosomes for each haplotype in the sample, k is the number of haplotypes, and p_i is the sample frequency of the *i-th* haplotype.

Both populations were found to be haplotypically diverse (haplotype diversity = 0.99 +/- 0.002 in Masalit; 0.99+/- 0.001 in Hausa). Haplotypes were found to be highly diverse across the region. Only a few haplotypes were found to be common (frequency > 5%) in the populations studied. Only two haplotypes in each population were found to have a frequency of more than 5%.

Using the software package ARLEQUIN (Schnieder et al. 2000) to look for the haplotypes shared between the two populations, I found that only 4 haplotypes were shared between Hausa and Masalit.

3.5.7. The degree of genetic differentiation between the Hausa and Masalit samples

Having two populations provided the opportunity to make comparisons between population groups. These comparisons were carried out using Fst plus three methods of genetic clustering.

3.5.7.1. Fst

A standard measure of gene frequency variation among populations is Wright's Fixation Index statistic (Fst). It reveals differences between populations by calculating the reduction in heterozygosity that is expected after random mating between the two populations. Fst measures the inter-population diversity using the difference between the average observed and the total expected heterzygosity. It was calculated for each of the SNPs typed using the equation described by Cavalli-Sforza (Cavalli-Sforza, Menozzi et al. 1994; Garte 2003). F_{ST} was first calculated between the two populations at each individual SNP. All SNPs except for three have an Fst value less than 0.06. The maximum F_{ST} value was 0.15 (Figure 3.5.7.1).

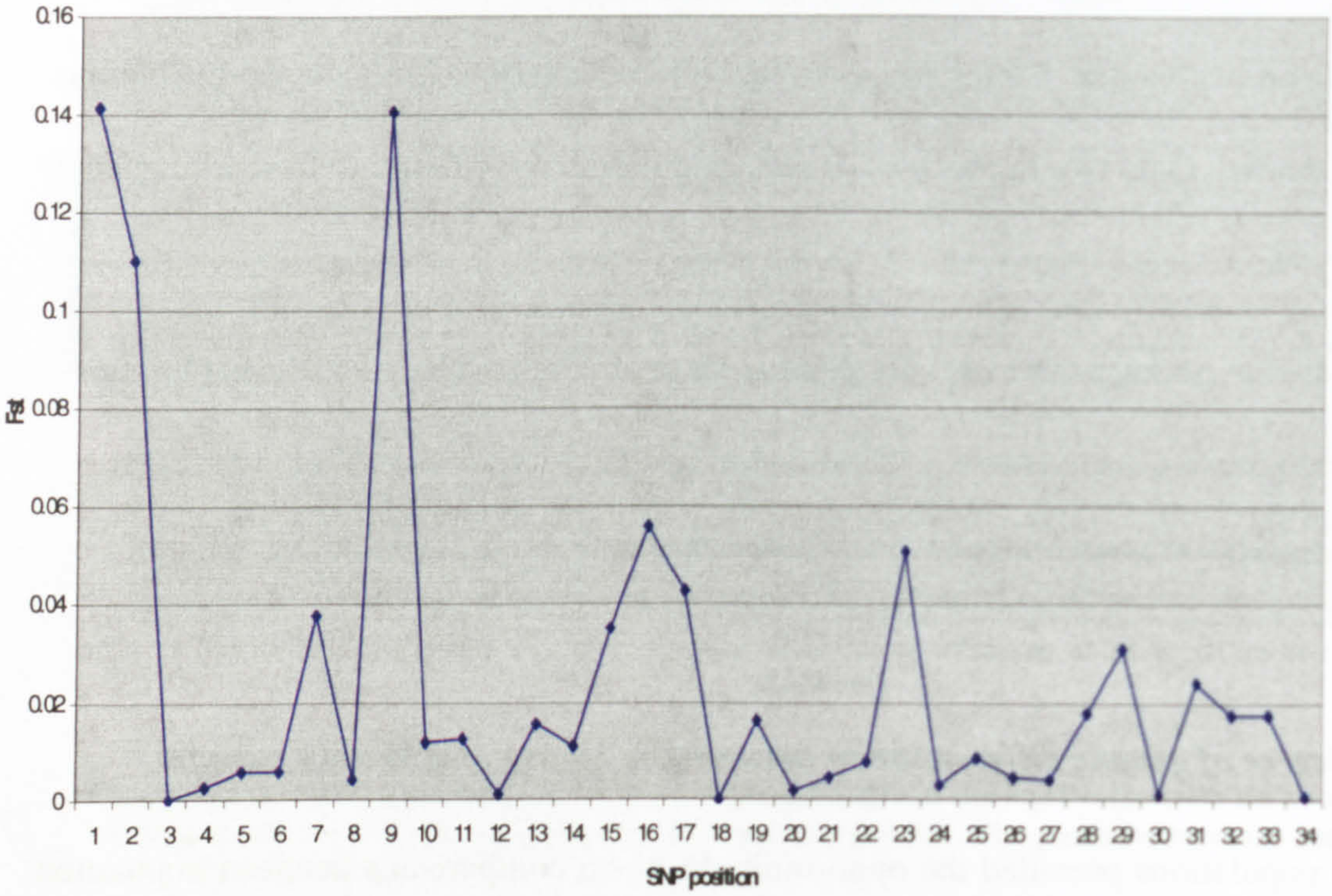


Figure 3.5.7.1: Fst values for SNPs typed in the 5q31 region in the Hausa and Masalit samples.
Markers ordered on the x axis as in Table 3.4.3. (See table for marker name and chromosomal position).

The fact that single markers' Fst values are low simply mean that the difference in allele frequency for each particular marker is not large enough for the test to be useful in drawing inferences about the genetic distance between the two populations (i.e. how genetically similar or different they are from each other). The average Fst value could be more significant, as it gives an overall estimate of the genetic distance across the whole area. This estimate is less influenced by the random allele frequency similarities of single markers. The mean F_{ST} value over all SNPs was 0.025, which is below the expected range of difference between two populations (0.07 to 0.15) (Rosenberg, Pritchard et al. 2002; Bamshad, Wooding et al. 2003).

Another approach I used to calculate the overall Fst value involves utilizing the full haplotypic information of all the typed markers, as is shown below.

Haplotypic Fst:

The ARLEQUIN software package was used to calculate the weighted average F-statistic over all loci to acquire a single general statistic that summarizes the difference between the two populations. The following formula was used:

$$Fst = \frac{f0 - f1}{1 - f1}$$

Where *f0 is the probability of identity by decent of two different haplotypes drawn from the same population, f1 is the probability by decent of two haplotypes drawn from different populations.*

The Fst value calculated using this method was 0.02 which was similar to the mean Fst values of single SNPs calculated previously.

3.5.7.2. Genetic Clustering Methods

1) ARLEQUIN:

The ARLEQUIN software package was used to assign individual genotypes to populations. This is done by determining the log-likelihood of each individual multi-locus genotype in each population, assuming that the individual comes from that population, taking into account the allele frequencies in each sample.

When the log-likelihood values of individuals from Hausa (pink dots) versus those from the Masalit tribe (blue dots) were plotted, as shown in figure 3.5.7.2a, it did not show two

distinct clusters. Rather, there was overlapping of the clusters, and a large number of individuals could have belonged to either group.

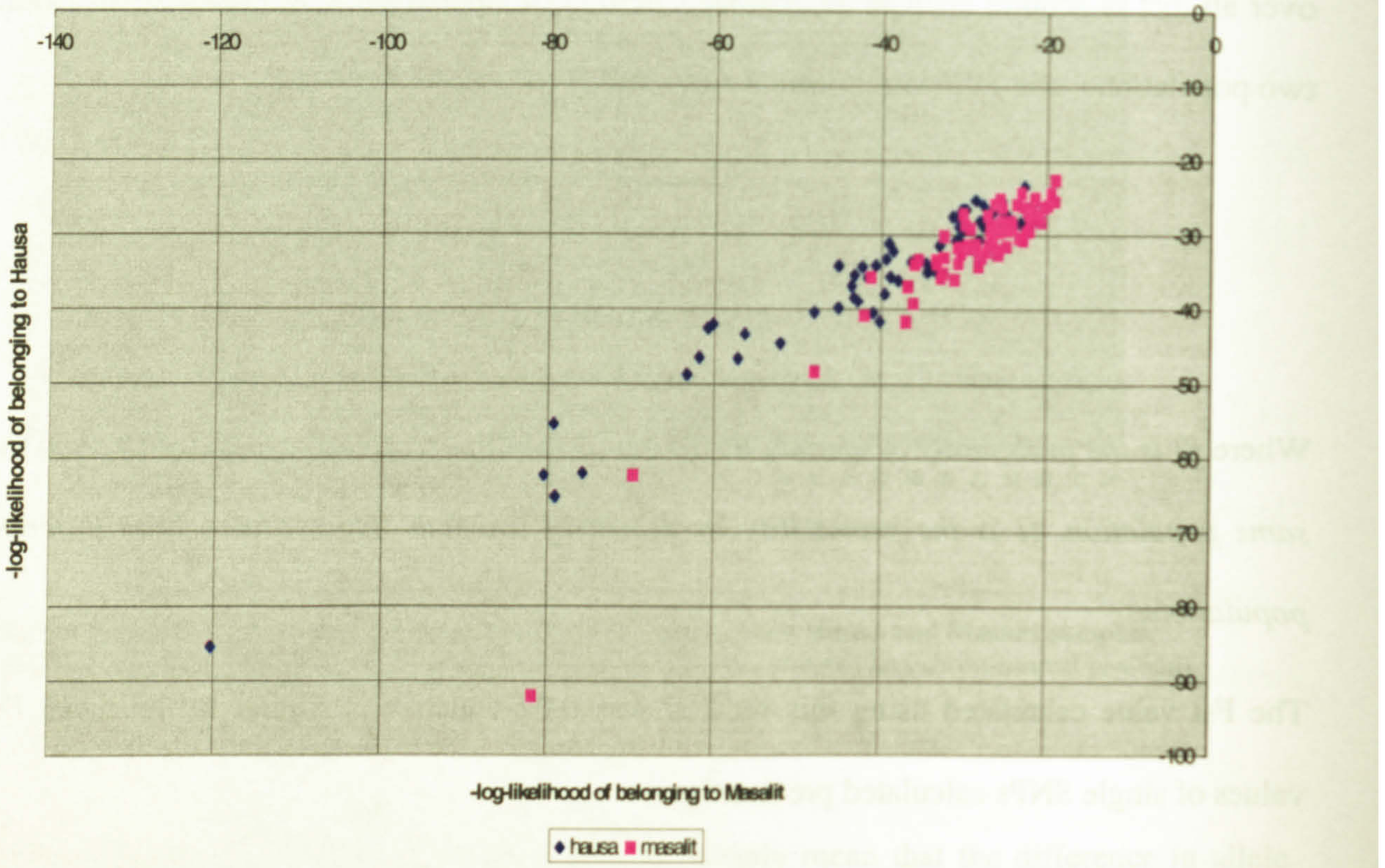


Figure 3.5.7.2a: Assigning individuals from the Hausa and Masalit to population groups by the Arlequin sofware. On the x-axis is plotted the –log likelihood of belonging to the Masalit. On the y-axis the –log likelihood of belonging to the Hausa is plotted. Each data point represents an individual.

2) Gene Genealogy:

The haplotypes from the Hausa and Masalit samples were pooled together after phasing the genotypic data in each group separately. A distance matrix was calculated for the pooled haplotypes. Then, as part of the software package PHYLIP version 3.5c (Felsenstein, J. 1993), the algorithm UPGMA was used to construct the phylogenetic gene tree that best describes the ancestral relationship between the haplotypes.

Because the pooled sample is constituted of two isolated groups, the alleles within each group are expected to be, on average, more similar to one another than comparisons between groups. This should produce two major clusters corresponding to the two populations in the constructed gene tree. The constructed gene tree did not have such distinct clusters corresponding to the Hausa and Masalit samples (figure 3.5.7.2b). The spacing of markers over a large genetic region, 29 markers over 650kb in this case, would have allowed some degree of recombination to play a part in shaping haplotype diversity, which consequently would have violated the assumptions of the tree building method.

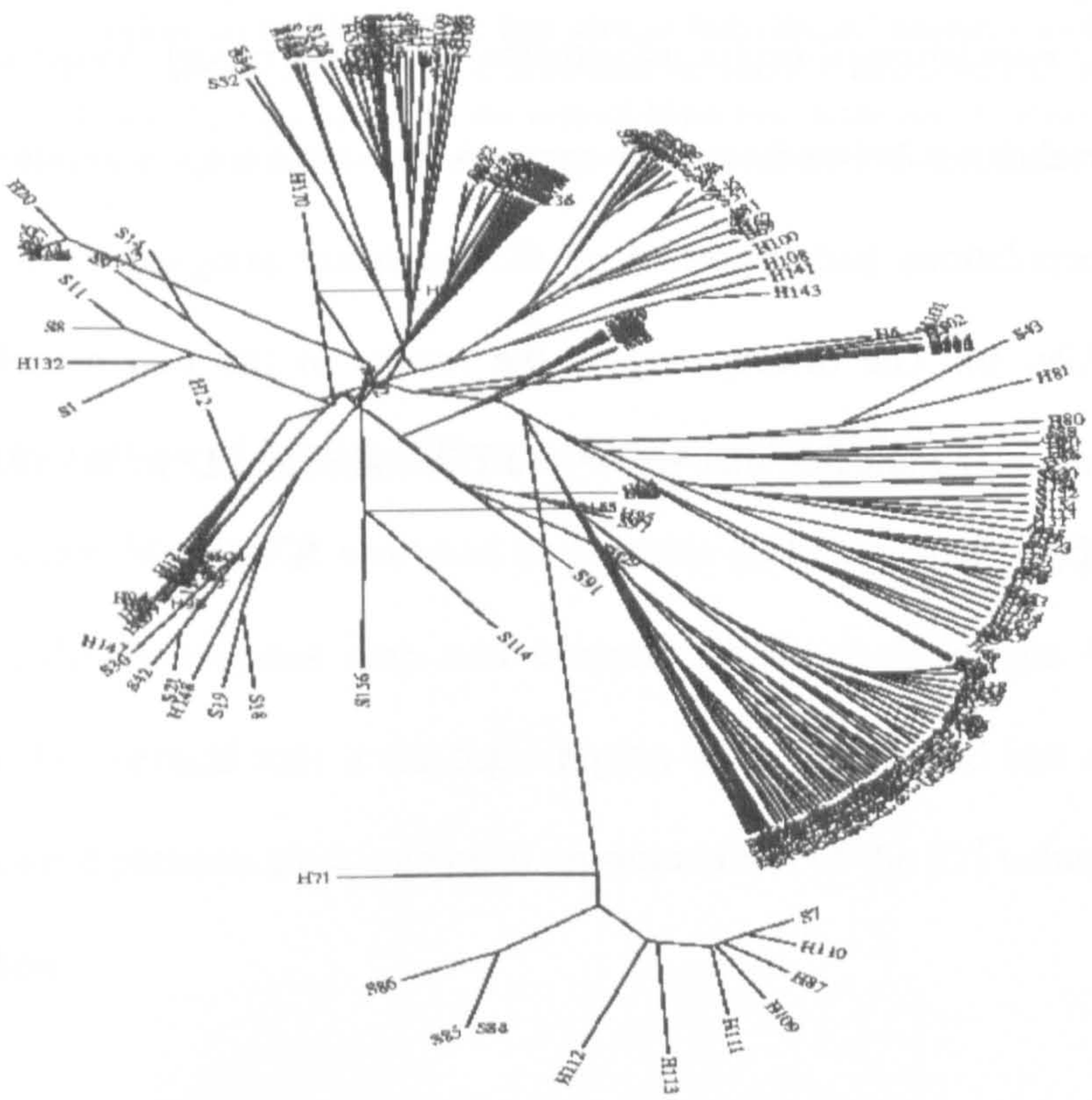


Figure 3.5.7.2b: Phylogenetic relationships between haploypes in the combined Hausa and Masalit sample. Shown is an unrooted tree with each branch representing a single haplotype.

3) STRUCTURE:

Using the program **STRUCTURE** 2.1 (Pritchard, Stephens et al. 2000), the two Sudanese populations were indistinct when the genotypes of the 29 polymorphic markers in the combined sample were run under a no-admixture model assuming two populations of origin. The model was provided with population-of-origin information for each individual. Each population had on average an equal proportion of its individuals assigned to one or the other population (Figure 3.5.7.2c, a). This could be either because there were not enough markers typed, or due to the presence of background LD between markers that is not accounted for by the program.

A linkage model was run using the haplotypes of the unrelated Hausa and Masalit individuals and providing population-of-origin information. With this model, *Structure* tries to account for the correlations between linked markers by assuming admixture. All individuals from both populations had a portion of their ancestry assigned to the other population (figure 3.5.7.2c, b). The discrepancy in the results of the two models (no-admixture and linkage models) indicates that there is a considerable LD in the data that impinges on the results of the analysis.

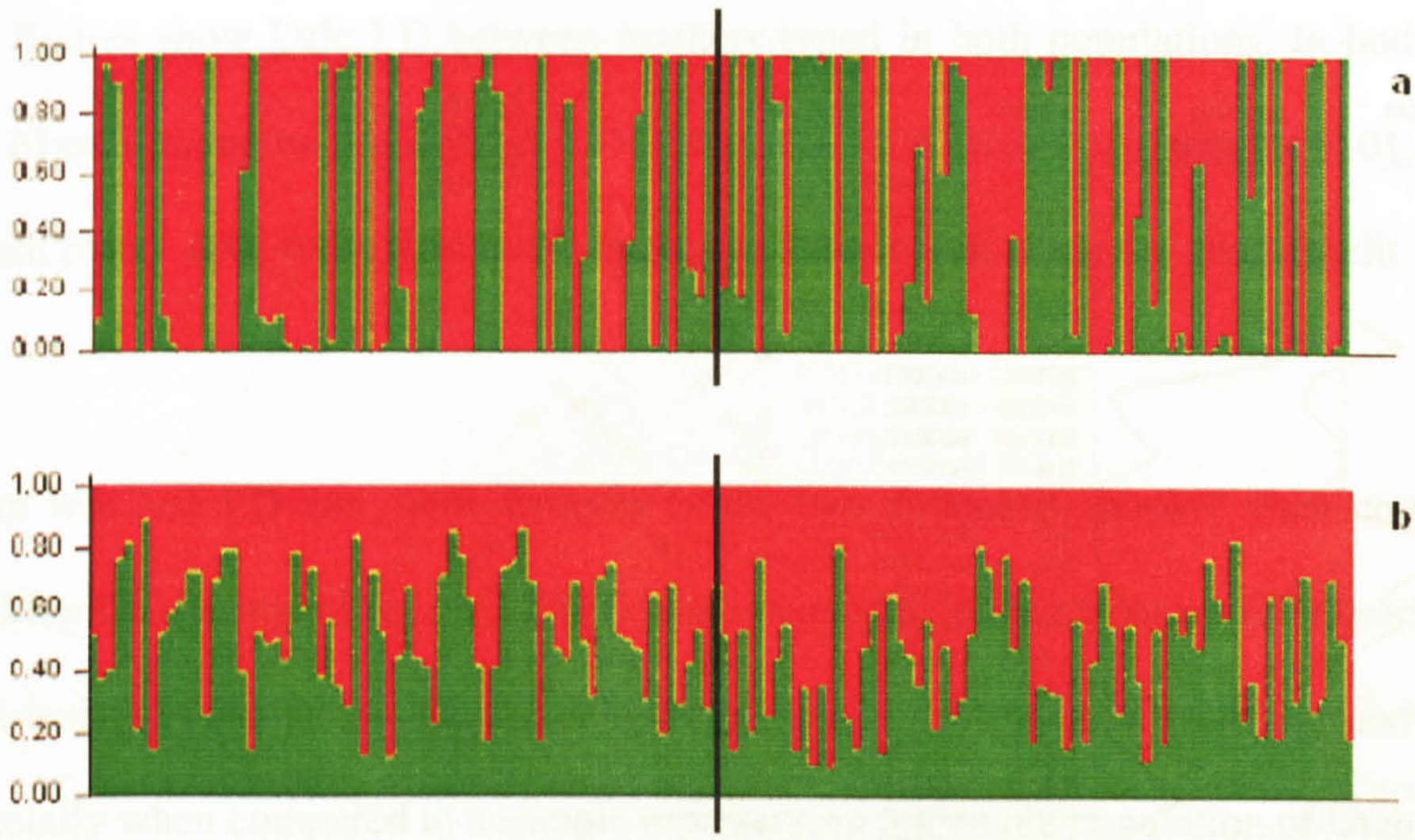


Figure 3.5.7.2c: a) STRUCTURE Bar plot of individuals' ancestry assuming no admixture and two populations of origin of the combined unrelated Hausa and Masalit samples typed for 29 markers in the 5q31 region. b) STRUCTURE Bar plot of individuals' ancestry under linkage model. In figure individuals are arranged on the x axis as vertical lines so that Hausa sample constitute left half of the graph and the Masalit the right half, with the vertical black line in the middle separating them. On the y axis the percentage of each individual's assigned ancestry is indicated with different colours for the two assumed populations.

3.5.8. The pattern of Linkage Disequilibrium in the 5q31

The program **MARKER** was used to generate an LD map of the 5q31 region in the Hausa and Masalit. A separate map was constructed for each sample (Figure 3.5.8a & figure 3.5.8b). The vertical axis is the 5q31 region with SNPs typed and minor allele frequencies, the coloured patterns are a statistical representation of the |D'| value calculated for each pair of markers.

Figure 3.5.8a: Marker Map illustrating the LD between SNPs in the 5q31 region in the Masalit. LD is measured by absolute D' (http://www.gmap.net/marker). Coloured spots connecting SNPs illustrate the absolute D' level between those SNPs. Colour coding is presented in the top right-hand corner.

Figure 3.5.8b: Marker Map illustrating the LD between SNPs in the 5q31 region in the Hausa. LD is measured by absolute D' (http://www.gmap.net/marker). Coloured spots connecting SNPs illustrate the absolute D' level between those SNPs. Colour coding is presented in the top right-hand corner.

The figures show little LD between markers typed in both populations. In both the Hausa and Masalit samples the average LD value was 0.05 with a variance of 0.01. LD values ranged from 1.8E-05 to 0.96 in the Hausa, and from 1.0E-05 to 1 in the Masalit.

There was less LD and more diversity in the two Sudanese samples when compared with HapMap CEU sample (Figure 3.5.8c). This is contrary to the extensive LD expected in small semi-isolated populations due to bottle necks, small effective population size and inbreeding, especially when compared to a sample representing the whole population of Utah.

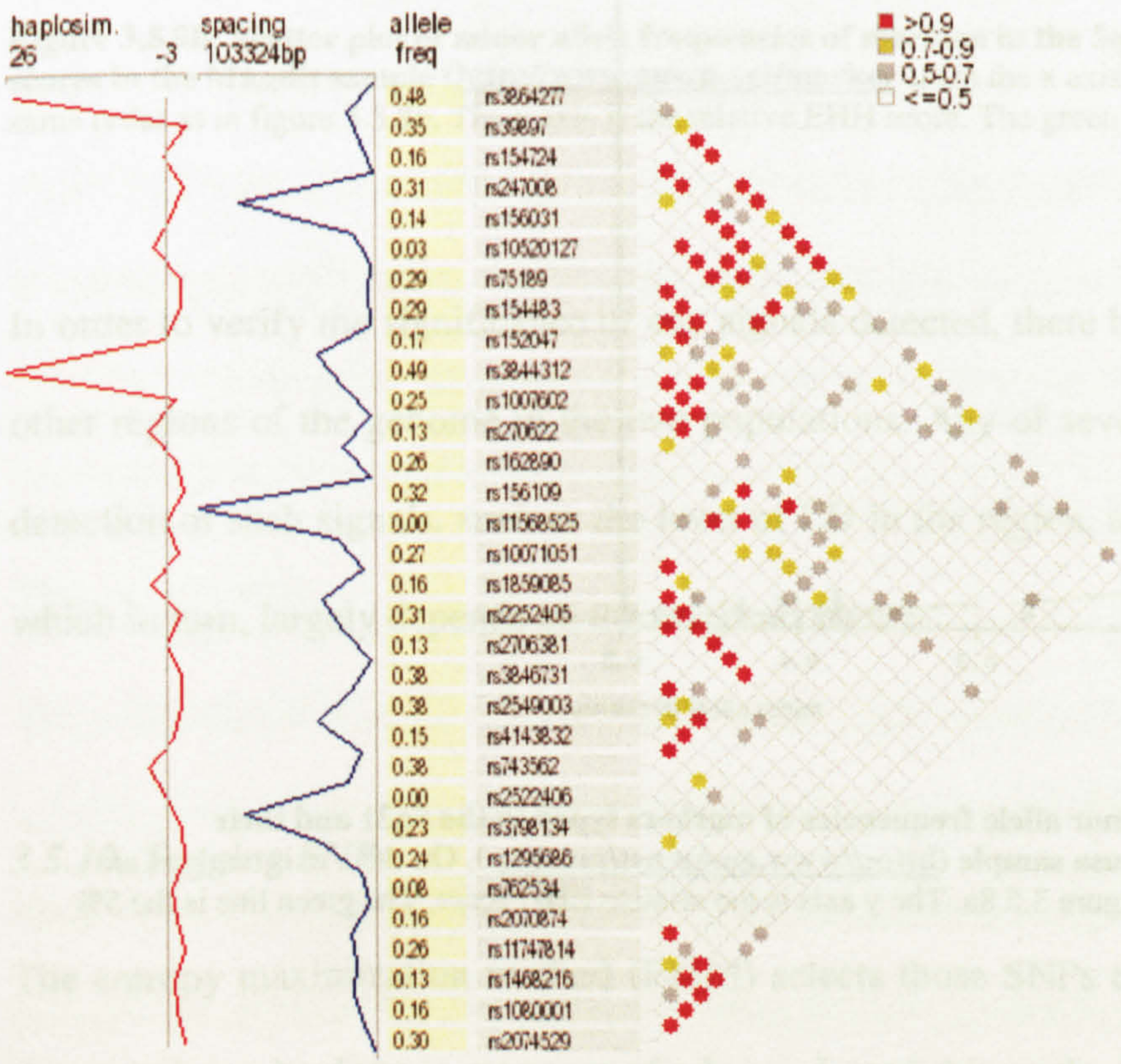


Figure 3.5.8c: Marker Map illustrating the LD between SNPs in the 5q31 region in the HapMap CEU population. LD is measured by absolute D' (http://www.gmap.net/marker). Coloured spots connecting SNPs illustrate the absolute D' level between those SNPs. Colour coding is presented in the top right-hand corner.

3.5.9. Signals of positive selection

To search for signals of positive selection in the 5q31 region in the two Sudanese populations, the haplosimilarity test implemented in the **MARKER** application (http://www.gmap.net/marker) was used.

There was no clear signal in the Hausa data (Figure 3.5.9a). The Masalit data revealed a very slight signal that was barely distinguishable from the background noise in the region (Figure 3.5.9b).

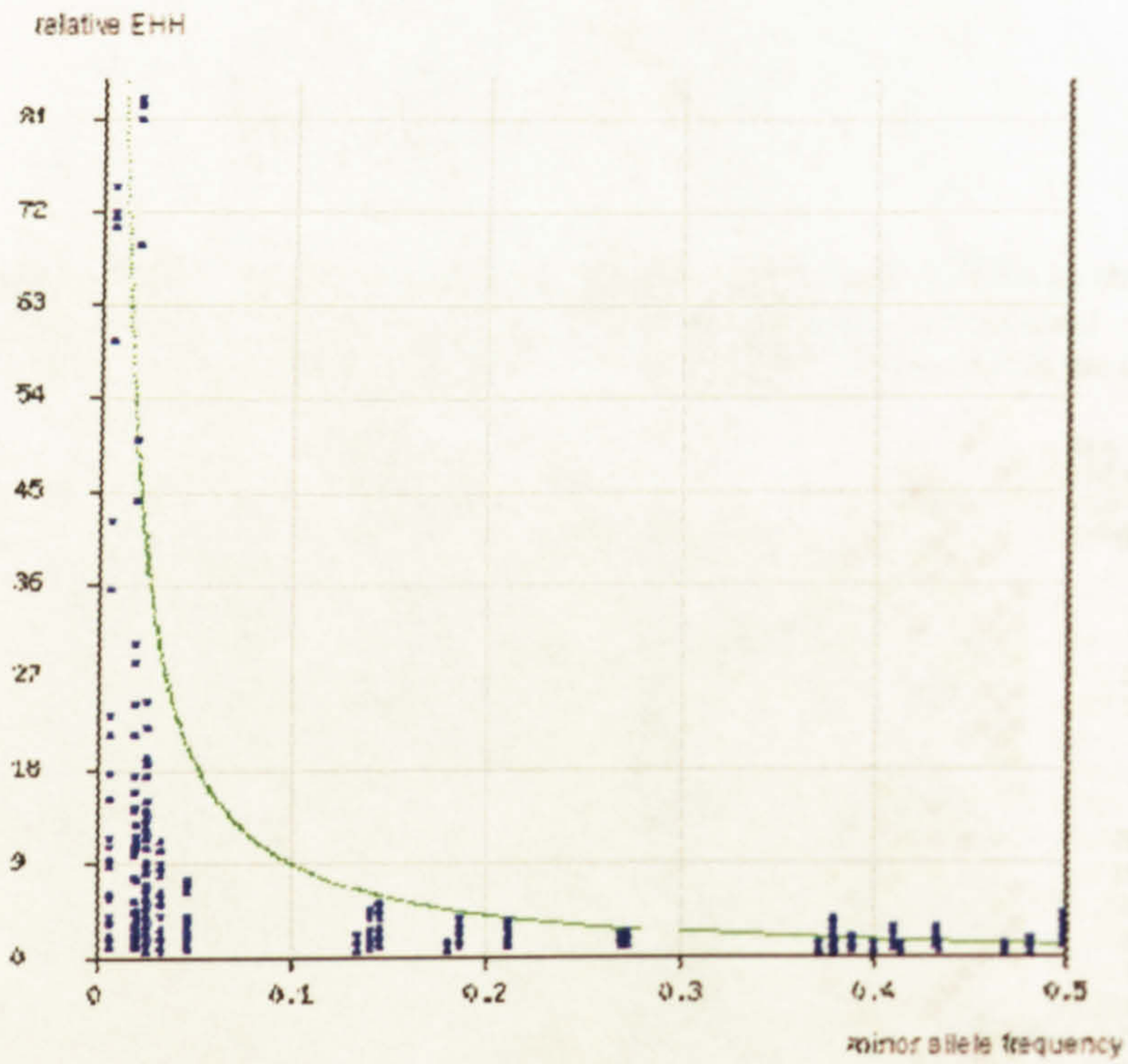


Figure 3.5.9a: Scatter plot of minor allele frequencies of markers typed in the 5q31 and their haplosimilarity scores in the Hausa sample (http://www.gmap.net/marker). On the x axis markers are arranged in the same order as in figure 3.5.8a. The y axis is the relative EHH score. The green line is the 5% significance level.

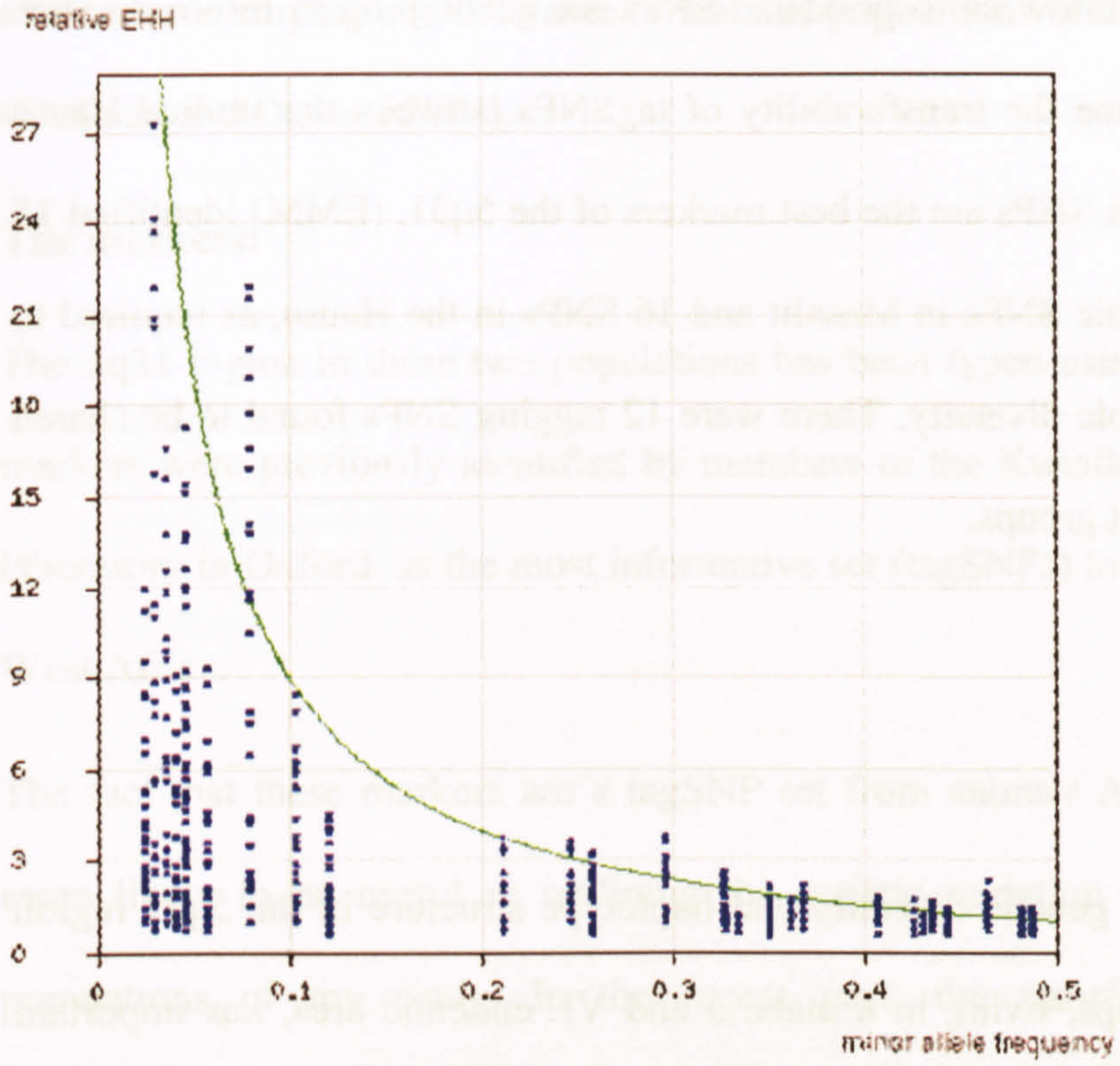


Figure 3.5.9b: Scatter plot of minor allele frequencies of markers in the 5q31 and their haplosimilarity scores in the Masalit sample (http://www.gmap.net/marker). On the x axis markers are arranged in the same order as in figure 3.5.8b. The y axis is the relative EHH score. The green line is the 5% significance level.

In order to verify the significance of any signals detected, there has to be a comparison with other regions of the genome in the two populations. Any of several factors could affect the detection of such signals, such as the level of LD in the region, and the haplotype inference, which in turn, largely depends on the markers' choice.

3.5.10. Tagging SNPs

The entropy maximization method (EMM) selects those SNPs that most effectively dissect the underlying haplotypic structure of a locus. I used this method to define a subset of SNPs that represent the greatest proportion of the full 29-SNP haplotypic diversity.

The results of these analyses allow us to prioritize SNPs for genotyping in future disease-association studies and examine the transferability of tagSNPs between the studied Hausa and Masalit. To identify which SNPs are the best markers of the 5q31, (EMM) identified 15 out of the 29 typed polymorphic SNPs in Masalit and 16 SNPs in the Hausa, as required to describe 100% of the haplotypic diversity. There were 12 tagging SNPs found to be shared between the Hausa and Masalit groups.

3.6. Discussion

Knowledge of the patterns of genetic diversity and haplotype structure in the 5q31 region across the two Sudanese groups, living in a malaria and VL endemic area, has important implications for identifying SNPs and haplotypes useful for genetic-mapping studies of these diseases in the two populations. Furthermore, this investigation offers an in-depth look at the genetic variation patterns in a typical African village setting. This knowledge especially when compared with that of populations from other parts of the world might help advance our understanding of human genetic diversity as a whole, in Africa and elsewhere.

Why the Hausa and Masalit are interesting:

The Hausa and Masalit populations of eastern Sudan were sampled from two neighboring villages with similar environmental pressures. They have distinct ethnicities and demography as evidenced by their history, how they identify themselves and their oral account of origin, migration and lack of interbreeding with other populations, as well as their linguistic differences. Founder effects, consanguinity, endogamy, and isolation from original population sources, might have played some part in creating homogeneity and in limiting variation in the genetic pools in these recently immigrant populations. Previously many

studies have emphasized the value of isolated populations in the dissection of complex traits (Shifman and Darvasi 2001).

The markers:

The 5q31 region in these two populations has been typed using a set of 29 markers. These markers were previously identified by members of the Kwiatkowski group, in the WTCHG laboratory in Oxford, as the most informative set (tagSNPs) in a sample from the Gambia in West Africa.

The fact that these markers are a tagSNP set from another African population made them more likely to be useful in outlining the genetic variation pattern in the two Sudanese populations of my study. In the recent past, the transferability of tagSNPs across populations, especially those from the same continental region, was suggested by several studies (Gu, Pakstis et al. 2007). And although Gonzalez-Neira et al. (Gonzalez-Neira, Ke et al. 2006) found Africa to be the most diverse region for the portability of tagSNPs from one population to another, nonetheless, they still found tagSNPs to be highly portable between African populations. However, those results were obtained in a gene-free region and may not be extended to other regions with different properties, like the 5q31 region.

The sample size:

The robustness of inferred LD structures is dependent, among other things, on the size of the samples used. Evaluations of empirical data recently concluded that 90-100 subjects will likely be sufficient for construction of stable enough HapMap (Zeggini, Rayner et al. 2005). Fallin et al. (Fallin and Schork 2000) estimated that a sample size of 100 is sufficient to estimate, using the EM algorithm, haplotype frequencies with a high degree of accuracy. For these reasons, I expected the LD and haplotype frequency estimations in my datasets to be

reasonably stable, especially with the advantage of typing extra family members, which provided more accurate haplotype phase information.

Polymorphism patterns:

This genotyping endeavour was carried out to determine the population patterns of genetic diversity in the Hausa and Masalit, their haplotypic and LD structure, and to gain insight into how genetically different these two populations are from each other in the 5q31 region. This particular region of the genome is functionally significant in many diseases. It harbours a number of key genes important in resistance to infections and in modulating the body immune response.

Although there are differences between the Hausa and Masalit of the study in terms of nutritional status, social organization and family structure, nonetheless, the fact that these two populations share the same environment and are exposed to the same pathogens, mainly *P. falciparum* and *L. donovani*; makes any patterns of genetic similarities above and beyond those expected, of particular interest, for it could indicate evolutionary convergence due to same selection forces acting on these two groups. This type of information could be potentially useful to aid in the design of the future association studies in these populations when phenotypes are considered, with the hope of understanding a bit more about the genetic factors underlying the susceptibility to these infectious diseases.

It became evident when I constructed the whole village pedigree that there was a high degree of relatedness between individuals from different families within the same village. The whole village could be divided into several clusters where there are many strong ties between families. From a recent study carried out in the Masalit – the same group studied here- whole genome scan data showed there are only a limited number of Y chromosomal lineages in Salala village (Miller, Fadl et al. 2007).

Although this setting – extended pedigrees with high degree of relatedness between them- might be ideal for some study designs like linkage studies and Family Based Association Testing (FBAT), it could present some challenges for others. For example it could potentially confound the results of case control studies, especially in founder populations that have grown rapidly and recently from a small size –as probably is the case here-, where there would be an increased likelihood of sampling bias toward collecting relatives (Voight and Pritchard 2005). The above underlines the importance of careful consideration of schemes adopted when sampling from groups with high degrees of relatedness between their members.

Surprisingly the expected genetic telltale signs of inbreeding and bottle necks, like increased homozygosity, the presence of a few haplotypes, and extensive LD; are all absent in the Hausa and Masalit samples, despite the fact that they both have a small population size and being founded by a few individuals who migrated to eastern Sudan and maintained a limited flux with their original population source. This can perhaps be partly explained by the fact that the rigorous criteria used for including individuals in the study might have biased the sample to include a disproportionate number of individuals from outside the village, which might have given rise to the picture of low LD in the region and lack of evidence of inbreeding. Alternatively low LD might arise due to exponential growth and rapid expansion of the size of these populations. The founding population for each village comprised 10–15 related males with their families. By 2004 population sizes were 1,300 in Salala village and 1,500 in Koka village. This rapid growth is concordant with the reported country-wide trend. In addition to being evidenced by the actual village numbers at present, the effect of population growth on patterns of genetic variation could also be magnified by the high mortality rate, which constitutes a sort of undetected or cryptic numbers.

The spacing and choice of markers from tagging SNPs which by definition have no or little LD between them might have made it more likely to get this result. Furthermore, the low LD observed in the 5q31 region could be related to the important functionality of the region, being packed with genes involved in many aspects of immunity to a wide range of diseases. It is possible that multiple forces of selection could be playing a role in shaping the polymorphism pattern in the region, resulting in a high preponderance of intermediate frequency alleles and low LD between variants. Fine-scale genetic map estimates from phase 2 HapMap data found genes involved in defence and immunity to have the highest recombination rates compared to genes of other functional classes (Frazer, Ballinger et al. 2007).

The amount of LD was very similar in the Hausa and Masalit samples. Although I have only quantified the amount of LD here, the question of quantifying the LD pattern differences between population groups remains an important one. I will be tackling this issue in the coming chapter. Just one example of the many arguments for its importance is that quantitative measures of transferability -of the set of tagSNPs selected from one population to another- are related to the extent of agreement (similarity) between the LD structures in different populations.

It was interesting to find less LD and more diversity in the two Sudanese populations when compared with HapMap CEU population. Especially considering the fact that each of the Sudanese groups represents a single, ethnically homogeneous small village, while HapMap CEU, on the other hand, was sampled from the whole population of Utah. The genetic diversity pattern of LD structure found in the 5q31 region in the two Sudanese populations versus that of the CEU population can be explained by the 'out-of-Africa' hypothesis of human dispersal. The subsequent founder effect in non-African populations and the larger effective population size of African population resulted in more genetic diversity, less LD

and higher heterogeneity among populations in Africa than elsewhere. However, these patterns could have been affected by other processes such as ascertainment bias of the markers, sample choice, or selective pressures.

Population differentiation:

The sampled Hausa and Masalit were similar in their minor allele frequencies with a tendency for Hausa to have higher MAF. All tests run to try and differentiate these two populations from each other (single marker Fst, haplotypic Fst, clustering individuals using ARLEQUIN, **STRUCTURE** and genetic trees.) failed to identify the two populations as genetically distinct from each other, and to cluster individuals correctly.

The F_{ST} values for individual SNPs confirmed the initial observation regarding allele frequency; namely, that the frequencies are relatively similar. The whole locus Fst, using haplotypes of all the SNPs I typed in the region, was 0.02. The mean single-SNP F_{ST} value was 0.025.

In a study conducted in 37 world populations using 80 independent loci, considerable substructure was found within geographical regions. More divergent genetic lineages and higher levels of subdivision have been shown to exist in African populations than in those from other regions. Estimates of average Fst values within Africa were found to be around 0.051 (Tishkoff and Verrelli 2003). In another study, in which 44 worldwide populations were typed at the CTLA4 gene, the highest continental Fst value was found in sub-Saharan Africa. The average Fst for African populations was 0.068 (Ramirez-Soriano, Lao et al. 2005). Kidd et al. (Kidd, Pakstis et al. 2004) found the distribution of Fst values in 38 world populations analysed with more than 100 neutral polymorphisms per analysis, to range between 0.042 and 0.380.

Since African populations are generally thought to have more associated diversity and to be more different from each other than would be predicted from their geographical proximity

(Kidd, Pakstis et al. 2004), the apparently minimal overall difference in diversity between the two Sudanese populations of my study was intriguing. It was also interesting to note that most of the SNPs with similar population frequencies in the Hausa and Masalit and therefore low Fst were of a high frequency. Kidd et al. (2004) noted that markers with low Fst values tend to be of low heterozygosities, and they found the combination of high heterozygosity and low Fst to be unusual. Whether this relatively low Fst in the 5q31 region is reflected in the genome as a whole, is an interesting question that merits further investigation.

Recent common ancestry can not be put forward as an explanation of the genetic similarities between the Hausa and Masalit, for they are West African and East African in origin respectively, and although they are geographically contiguous at present, high levels of migration and gene flow can not explain their similarities either. They are both highly endogamous with marriages exclusively restricted to within the same ethnic group. Furthermore, what makes the above two possibility even less likely is the wide social and linguistic divide of the sampled Hausa and Masalit.

Determining the F_{ST} over several SNPs at a particular locus (average F_{ST}) is suggestive of the forces that might have produced the particular F_{ST} distribution. Fst values that are exceptionally high or low could reflect differential selection acting on particular loci rather than genetic drift or migration. For example, a low global F_{ST} could indicate balancing selection where the allele frequencies are expected to be similar, while high global F_{ST} values could be the result of directional selection where there is divergence of allele frequencies (Hamblin and Di Rienzo 2000). Due to the importance of the 5q31 region in modulating the immune response to infection, and the commonality of the environmental pressures in the two villages; this genomic area could have been shaped by balancing selection acting on the two populations and creating an excess of intermediate-allele-

frequency markers. This raises the need to further explore the signals of positive selection in the region.

But the most likely explanation could be that this observed pattern is a consequence of the density, spacing and choice of markers. It could very well be that the number and characteristics of typed markers, does not allow for enough resolution to distinguish these two populations from each other. Marker choice affects downstream statistics based on the allele frequencies of these markers, like Fst and algorithms used in **STRUCTURE** and ARLEQUIN. For methods that rely on haplotype information like gene genealogy inference, robustness of the phasing of the extended haplotypes over a 600 kb region using a few markers might be a consideration.

Although it is difficult to advance any definitive answers for these observations, in the next chapter I will explore the markers-set choice further, and try and maximize markers' information content used to tease out the genetic distinctness of the Hausa and Masalit. This will be done to examine whether there is enough between-population variance, sufficient to cause consistent bias in case control studies when such subtle differences can be made significant by the larger scale sampling schemes typically employed by these studies.

Signals of selection:

Most SNP data have been obtained by choosing high frequency markers from publicly available databases. The process by which the SNPs have been selected affects levels of LD observed in the data and the frequency spectrum, which makes these studies not ideally suited for detecting selection. Ascertainment bias complicates downstream analyses of selection signals, one example of that might be the skew it creates in allele frequency

spectrum. That is why for this dataset the metric I've chosen to look for signals of selection depends on the haplotypic information rather than allele frequencies.

While the Hausa did not display any obvious signal of selection in the 5q31 region using the Haplosimilarity measure, the Masalit had a small signal that did not stand out against other background signals. This emphasizes the difficulties in weighing the significance of such results without a standard to measure against. Understanding how frequently signals of this magnitude can occur by chance alone, as well as interpreting them against background noise, is important in identifying markers subject to positive selection.

Patterns of genetic variation and LD are affected both by gene specific factors (mutation, recombination rate, conversion, selection) as well as demographic histories (contractions, expansions, subdivision). In contrast to population processes that affect the whole genome, gene factors such as differential selection can shape the haplotype structure and LD in specific gene regions and might result in population differences. Since the 5q31 region has a key role in the immune system and has been related to susceptibilities to infectious diseases, the exposure to geographic differential selective pressures, such as the presence of pathogens that might have affected the 5q31 gene structure, could be envisaged. This has been the case of the selection for resistance to malaria detected in several genes such as G6PD, Duffy and TNFSF5. But demographic explanations have to be ruled out first before selection can be put forward as an explanation.

It would be interesting to study these two populations at another region in the genome where selection might have played a part and where the functional variant is already known, and to compare the selection signal with the patterns seen here at the 5q31 region. With the large body of evidence of it being under malaria selection, the HbS locus is one such region

perfect for modeling how the signal of positive selection would look like in these two populations.

Furthermore, any additional genotyping carried out in another genomic region could help improve the resolution of genotypic data to distinguish the two populations as separate entities.

Impact on association studies:

The knowledge of the haplotype structure and LD patterns in specific regions, such as the 5q31 region, will shed light not only on the history of the populations analysed but also on the genomic processes that could be pivotal for biomedical interests.

Although there were striking similarities in allele frequencies and amount of LD, and less than expected structuring by available genetic distance estimates; significant differences in haplotype composition were found to exist between the geographically contiguous Hausa and Masalit of eastern Sudan. Genomic processes might have affected the populations in the same manner, giving similar LD and allele frequency patterns, whereas the differences found in haplotypes, might be explained by demographic processes, such as expansions, founder effects and migrations.

Although the differences observed here are small, they could be highly significant in the context of a large case-control study in which ethnic groups were not well matched between cases and controls. These results point to the need of a very well matched control population to compare with cases in order to minimize false positive associations.

The analysis done on the 5q31 region could be linked to improving the future design of any association study to be carried out in these two Sudanese populations. Prioritization of markers to be genotyped could be done by choosing the markers with the greatest chance of identifying those contributing to the phenotype under investigation like typing the tagSNPs

or those with some evidence of being under selection; thus reducing financial and statistical costs of studies.

These data could have several additional roles in the analysis of disease-association studies. For example, using phased haplotypes with high probabilities from this small set of trios could help in phasing the haplotypes of a larger case control set of unrelated individuals from the same populations.

By a similar argument, missing genotypes out of the set used here, either because of genotyping failure or because the SNP was not assayed in the case control study, could potentially be inferred through comparison to the genotyping data generated in this analysis.

3.7. Conclusion

A map of LD and haplotype structure in the 5q31 region has been constructed for each of the two Sudanese populations separately. The two populations were found to be very similar in terms of their minor allele frequencies and genetic distance. On the other hand, there was negligible overlap in the haplotype frequency between the two groups. There was also little LD between this particular set of markers in the Sudanese populations when compared with a population of European ancestry. It was also difficult to interpret the positive selection signals of this dataset in relation to background noise.

Sampling, choice of markers, demographic factors or selective forces can all influence the results, and consequently would affect any downstream statistics. This exploration of the data has certainly raised more questions about African-population genetics than answers. The rest of the thesis will be an in-depth exploration of these questions, and the practical applications of the knowledge gained through these explorations.

As all the available methods used to discern the genetic differentiation between the Hausa and Masalit failed to reflect this difference in my dataset. Bearing the limitation of the data firmly in mind, I will go on, in the next chapter, to explore an LD-based statistical analysis that will maximize the informativeness of markers for estimating genetic distance.

In chapter 5 I examine the β-globin region harbouring the sickle haemoglobin variant, as a clear example where natural selection is known to have played a part in shaping its diversity and where the functional variant is known, to allow an easier interpretation of the genetic variation patterns associated with positive selective pressures in the Hausa and Masalit.

Chapter 4:

Ascertaining genetic differentiation between closely related populations by employing LD information of a limited set of linked markers.

4.1. Abstract

In the previous chapter, methods employed to genetically differentiate the ethnically distinct Hausa and Masalit did not yield significant results. However, there were marked differences in allele association patterns and haplotypic structure, in spite of the observed allele frequency similarities. The focus of this chapter is determining whether this apparent difference is a product of chance, or whether it is a reflection of diverse ancestry, and if so, can it be usefully employed in genetic distance estimation. In order to answer these questions I wrote and developed programming scripts to calculate pair-wise LD values for markers across the genomic region for each population group separately using the Expectation Maximization (EM) algorithm. Then the correlation between the paired samples was calculated using Spearman's rank correlation coefficient (rho). To test for the significance of the results, I employed a permutation approach. Between a 1000 and 50000 bootstrap samples were generated for each pair of populations compared.

I performed this analysis on limited sets of linked markers, along a 650 kb stretch of chromosome 5q31 region, in four African population samples, comprising the Hausa and Masalit of eastern Sudan, a population sample from the Gambia, and the HapMap Yoruba sample from Ibadan, Nigeria. I also compared those groups with the HapMap CEU sample of Utah residents of European decent.

Incorporating LD information proved successful in highlighting the genetic divergence of Hausa and Masalit and for most of the between-African-populations comparisons, with a P value as low as 0.0008. When the HapMap CEU population was compared with the African groups, there was more than forty fold decrease in the P value.

4.2. Objectives

- Explore whether LD pattern differences correlate with genetic distances between distinct population groups.
- Employ a permutation approach to test for statistical significance.
- Compare the results against other available methods for determining genetic differentiation.

4.3. Introduction

Characterization and quantification of genetic diversity has long been a major goal in evolutionary biology. As well as the relevance of this topic to understanding human population history and anthropology, it is of importance for the investigation of genes associated with disease. The idea is that members of a preconceived ethnic group share common ancestry that may include genetic risk factors. Human variation has been shaped by the long-term processes of population history, and population samples that reflect that history carry statistical information about shared genetic variation or ancestry.

As well as lending itself to the analysis of case control association studies - by highlighting any hidden population structure in the sample that might generate spurious results if undetected- discerning genetic differentiation between populations might be hugely beneficial in the field of pharmacogenetics, where the genetic structure of a population is

used as a predictor of the efficacy of drugs or the likelihood of adverse reactions (Jorde and Wooding 2004).

Populations are often defined in many often arbitrary ways. Exploring human differentiation could be pursued using frameworks that incorporate information on geography, culture, language, ethnicity and local demography, as proxies for the genetic makeup. Genetic distance between populations was found to increase with geographic distance at both continental and global scales (Cavalli-Sforza and Feldman 2003). Although studies revealed that geographic distance explains at least 75% of the variance between human populations (Manica, Prugnolle et al. 2005), still, it is not enough to explain the whole of the genetic variation. The correlation between the results of genetic analysis and concepts of race, language or ancestry might not be perfect, therefore direct assessment of genetic variation will yield more beneficial information, especially if distance between groups is described in a context relevant manner, like accounting for LD when designing and analysing association studies (Jorde and Wooding 2004).

Patterns of LD are the product of population histories and human migrations, recent natural selection, and the distribution and evolution of recombination hotspots. Previous studies of LD patterns in the human genome have shown that LD appeared to vary substantially among populations and is sensitive to the demographic history of a population, like founder events, bottlenecks and isolation (Slatkin 1994; Service, DeYoung et al. 2006). LD can extend over large genetic distances in isolated population groups, which was found to be quite different from its extent in outbred populations, in spite of the remarkably similar heterozygosities (Angius, Hyland et al. 2008). Even neighbouring isolate villages were found to be different in their genetic background (Angius, Bebbere et al. 2002).

LD serves as the backdrop for the design of association mapping studies (Zondervan and Cardon 2004). Based on this fact, it makes sense to look at any substructuring in the sample using a method that employs LD information. As it is often the case that the risk-enhancing SNPs are not observed directly, association tests consequently rely on LD between the observed and unobserved causal variant. And it is the pattern and extent of LD that determines the feasibility and design of association studies. This dependence on LD becomes even more pronounced when haplotypes are used to test associations. Testing for LD pattern homogeneity between groups making up the sample goes a longer way towards minimizing type 1 error than relying on allele frequency information alone.

Furthermore, testing for LD pattern conservation across populations is of paramount importance in the prediction of successful transferability and efficiency of tagging SNPs derived from one population to be used for an indirect association study in another population. Recently, some studies have revealed significant variation in the underlying haplotype structure in spite of the observed conservation of tagSNP patterns across global populations (Gu, Pakstis et al. 2007). This might indicate that even in cases where the coverage of tagSNPs appears to be preserved across populations, caution still needs to be exercised because the hidden genetic variants tagged by any particular tagSNP might not be the same in different populations.

The HapMap samples have been extensively used in designing studies and guiding analysis in other populations. Marchini et al. (Marchini, Howie et al. 2007) have successfully used the HapMap data to impute unobserved SNPs in the Wellcome Trust Case Control Consortium (WTCCC) samples. This might have been made more feasible because of the close genetic proximity between European populations. But for the more diverse relationships between other populations, initial testing of LD pattern comparability between those populations could be factored into the imputation of unobserved SNPs.

For most of the currently available genetic distance methods where the differences in allele frequencies between the populations is the corner stone; no account is taken of the LD relationships between markers, on the contrary, any strong LD in the data might present a challenge for interpreting the results of these analyses. The need for new measures of genetic distance has been highlighted for genotyping data with marked LD between marker loci. Employing an inappropriate genetic distance estimator can lead to misleading conclusions in this type of data (Pritchard, Stephens et al. 2000).

Some examples of the most frequently used genetic distance methods that rely on allele frequency data are: Nei's genetic distance (Nei and Feldman 1972), Fst and the derived Nei's coefficient of differentiation (Nei 1973). Because genetic diversity between humans is much less than those of many other species (Li and Sadler 1991) and they are found to vary only slightly when metrics like Fst and average nucleotide diversity(π) that depend on allele frequency comparisons are used, the accurate classification of this diversity is extremely sensitive to the way markers are ascertained. For example choice of markers may favour those of high frequency for their high information content, especially when tagging sets are employed across populations. These high frequency markers are more likely to be old and shared between populations, therefore using them might lead to an underestimation of the genetic distance between groups compared. Due to the absence of meaningful quantitative cut off points that can be applied across analyses, there also arises the difficulty of deciding what a particular result means. Therefore Fst and related measures are more suited to be used in a relative context when more than two population groups are compared, or when comparisons are made between different functional classes of genes (Kullo and Ding 2007). Metrics like Fst treat each locus in isolation from others and takes no account of correlations among loci. Consequently, the more loci analysed, the more of the LD information is ignored by this class of metrics.

Distance based clustering methods like the program **STRUCTURE** also uses allele frequency information. The model in **STRUCTURE** assumes that markers are not in linkage disequilibrium within subpopulations, so it can not handle markers that are extremely close together (Falush, Stephens et al. 2003). In addition, it usually requires the use of a large number of loci to tease out this difference and identify population structuring.

Differences in haplotype frequencies between the various compared groups could be considered to overcome the problem of LD in the data, like Rogers distance(R) (Rogers 1972), but using haplotypes across relatively big genetic regions would raise the issue of the extent of haplotypes and their constituting markers that are most appropriate to use. If too large a genomic segment or too many markers are considered, most haplotypes would be of a single or very low frequency, subsequently there might not be a difference in diversity within and between groups. i.e.; the chance of randomly picking two distinct haplotypes from different groups would be the same as sampling from the same group. On the other hand if shorter haplotypes are used, a lot of potentially useful information would not be utilized, which might lead to loss of power. Furthermore, because the accurate estimation of the haplotypic phase is paramount to this type of approach, errors in phase determination are a concern.

In this chapter, I explore the usefulness of comparing LD patterns across population samples and its implications for highlighting genetic divergence. I examine the question of whether taking account of LD differences might possibly provide a useful addition to genetic distance estimation methods, and be more sensitive in distinguishing more closely related groups by getting the most out of the data, especially in my datasets of limited numbers of markers in a restricted genomic region, typed in a few individuals. In spite of the recent accessibility of genome wide genotypic data, this kind of data in a restricted genomic region is still relevant and usually encountered in studies with gene targeted approaches, where a lot of markers are

typed within or around a gene of interest, or when sequence data is available for such genes or regions.

In addition to carrying out the analysis in the populations of my study that enjoy proximity by geographical distance, i.e. the Hausa and Masalit inhabiting two neighbouring villages in a remote area of eastern Sudan, I analysed another two African populations, the Gambians and YRI populations, that share a west African origin with the Hausa. I also compared the LD in those African groups with that of the HapMap CEU.

I used a limited number of polymorphic genetic markers (23-30 SNPs), with significant LD between them, typed in a short segment of the 5q31 region (about 650kb).

When compared with the currently available methods in ability to reflect the dissimilarities between these populations' genetic makeup, the LD-based approach showed consistent results with some methods, but it had higher resolution and discriminatory power for unravelling human population structure than most methods.

4.4. Materials and Methods

4.4.1. Population samples

For samples used in the study, proper informed consents and ethical approvals were obtained.

The Sudanese dataset:

In Sudan, sampling was carried out in two villages along the bank of Rahad river area of eastern Sudan: Koka village founded by the Hausa and Salala village founded by the Masalit (see Materials and Methods Chapter).

Family histories were reviewed for consanguinity and relatedness of the individuals. Only those trios with no kinship either within or between them were chosen.

From each village 96 individuals were genotyped: initially for the Masalit; 63 were in trios, 12 in parent & child pairs, and 21 were unrelated. For the Hausa: 45 were in trios, 32 in parent &child pairs, and 19 unrelated. For 11% of the sampled individuals, their pedigree did not concur with their genotypic data. For every trio with two or more markers with pedigree inconsistencies out of a possible 30, the three individuals making the trio were analysed as unrelated. After the pedigree check there was 72 unrelated Masalit and 72 unrelated Hausa

The Gambian Dataset:

The Gambian genotyping data was obtained with kind permission from the authors (Luoni, Forton et al. 2005). The sample comprised 128 unrelated Gambian chromosomes, based on genotyping 32 family trios.

The HapMap Dataset:

These analyses are based on release 22 (March 2007) of the genotype data generated by the International HapMap Consortium. Genotypes were retrieved from the HapMap Project Web site (http://www.hapmap.org) using HapMart for 24 markers in 30 CEPH trios from Utah (CEU sample, 120 independent founder chromosomes), and 23 markers in 30 Yoruba trios from Ibadan, Nigeria (YRI sample, 120 independent founder chromosomes).

4.4.2. Marker selection

The 30 single nucleotide polymorphisms (SNPs) set that was typed in the Sudanese populations was based on markers that have previously been tested in the laboratory in Oxford, in the Gambian sample. These markers were selected as the most efficient set of markers to capture most of the haplotypic diversity in the Gambian population (haplotype tagging SNPs), from a larger set of markers in the 5q31 region (Luoni, Forton et al. 2005).

4.4.3. SNP genotyping

For the Sudanese samples, DNA was collected using buccal brush and extraction was performed by a standard guanidine hydrochloride protocol at the Institute of Endemic Diseases, University of Khartoum. Total yield for a sample was 20 ug on average. DNA concentrations were measured and dilutions normalized to 20ng/ul using the picogreen DNA quantification kit. Whole genome amplification was performed using Primer Extension Pre-amplification PCR Method (PEP) at the Wellcome Trust Centre for Human Genetics, University of Oxford. Genotyping was carried out using the MALDI-TOF Mass Spectrometry (matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry) system (Sequenom). The proportion of missing data was not more than 3%. Missing genotypes were distributed across individuals from the two population groups.

4.4.4. Genomic region

The genomic area typed spanned a 646.5 kb in the 5q31 region (5:131443826-5:132090325).

The 5q31 region was selected because it contains several genes encoding molecules that have important functions in the regulation of the immune response, such as IL-4, IL-13, IL-5, IL-3, CSF and IRF1. It has been implicated in the susceptibility for parasitic infections like malaria and schistosomiasis.

The extent of genomic region and choice of markers to be typed as high frequency tagging SNPs is described elsewhere (Luoni, Forton et al. 2005). The same 650 kb segment of the 5q31 region studied in the Gambian population was chosen to be typed in the two Sudanese populations with the tagging SNPs identified from that study. And it is data for this SNP set that was downloaded from the HapMap website for the YRI and CEU populations.

4.4.5. Statistical analysis

All markers were tested for departures from Hardy-Weinberg equilibrium. The HWE test for each SNP within each population was calculated by standard χ^2 statistics, and none of the tested SNPs were found to deviate from HWE at a significance value of 0.01. To compare allele frequency between population groups; after calculating allele frequencies in each group of unrelated individuals, each marker was compared between population groups, using a 2x2 chi square test with one degree of freedom.

Using the software package **PHASE (v 2.1)** (Stephens, Smith et al. 2001; Stephens and Donnelly 2003) to infer the chromosomal phase of the parental genotypes, haplotypes for each population sample were generated by integrating family- and population-based reconstruction methods.

KOIND: Using the KOIND package (Kosman and Leonard 2007), several within-population diversity measures were calculated (Nei(Hs), Muller(Mu), Kosman expected(K), Simpson(Si)). Values close to 0 indicate high uniformity, while large values indicate high diversity. 200 bootstrap samples were generated for each population's haplotypes. Measures of diversity were averaged over all bootstrap-derived estimates. Consider a sample collected from population P, which consists of n haplotypes $x_1, x_2, \ldots, x_n$ typed at k bi-allelic loci. q_i denotes the frequency of allele1 at the ith locus, $i = 1, 2, \ldots, k$. The number of copies and frequency of haplotype r in population P are denoted by n_r and p_r, $r = 1, 2, \ldots, s$, respectively, where s is the number of distinct haplotypes in P. The measure of dissimilarity between haplotypes is denoted by ρ.

Kosman index	K	$K(P) = \sum \min[2qi,2(1-qi)]/k,\ 1 \le i \le k,\ 0 \le K(P) \le 1$
Müller index of diversity	Mu	$Mu(P) = [2n/(n-1)]\sum qi(1-qi)/k,\ 1 \le i \le k,\ 0 \le Mu(P) \le 1$
Nei measure of gene diversity	Hs	$Hs(P) = \sum [1-q_i^2-(1-q_i)^2]/k,\ 1 \le i \le k,\ 0 \le Hs(P) \le 0.5$
Simpson index	Si	$Si(P) = 1-\sum p_r^2,\ 1 \le r \le s,\ s \le Si(P) \le 1$

Between-populations diversity measures were calculated also calculated using the same package (The Nei coefficient of differentiation (Gst), The Kosman-Leonard expected, The Rogers distance(R)). Values close to 0 indicate very little genetic differentiation. 200 bootstrap samples were generated for each population's haplotypes. Measures of genetic distance were averaged over all bootstrap-derived estimates. Consider two samples collected from two populations P_1 and P_2, which consist of the same number n of haplotypes $x_{11}, x_{12}, \ldots, x_{1n}$ and $x_{21}, x_{22}, \ldots, x_{2n}$, respectively, typed at k bi-allelic loci. q_{1i} and q_{2i} denote the frequencies of allele1 at the ith locus for populations P_1 and P_2, respectively. The frequencies of haplotype r in populations P_1 and P_2 are denoted by p_{1r} and p_{2r}, respectively, $r = 1, 2, \ldots, s$, where s is the total number of distinct haplotypes in both populations. The measure of dissimilarity between haplotypes is denoted by ρ.

Rogers distance	R	$R(P1,P2) = \sum \lvert p1r-p2r \rvert/2,\ 1 \le r \le s$
Nei coefficient of differentiation	*G*st	$Gst(P1,P2) = 1/k \sum Gsti(P1,P2),\ 1 \le i \le k$. (Nei 1973)

***STRUCTURE v2.1*:** Analysis was carried out with 100,000 burning and 100,000 iterations.

Two models were used in the analysis: the no-admixture model, where the LD in the data is ignored, assuming two populations of origin. The model was provided with population-of-origin information for each individual. The other model is the linkage model, when any LD

in the data is attributed to admixture in the population history, the linkage model was run using the phased haplotypes of the unrelated individuals and providing population-of-origin information. For estimating K, 1,000,000 iterations were used for assumed number of populations (k) between 1 and 10.

Pairwise Linkage disequilibrium coefficients: Haplotype frequencies for all pairs of SNPs were estimated by maximum likelihood using an implementation of the EM (Expectation-Maximization) algorithm. Pair-wise disequilibrium was summarized using the r^2 measure, which was calculated using estimated haplotype frequencies. In each pair of populations compared, only the markers typed in both groups were used in the LD pattern comparison. For LD quantification and summary in all population groups, only the 23 markers shared between all populations were used.

Description of the LD-based approach: The theoretical hypothesis behind this analysis is that for the degree of correlation in the MAF between any compared groups belonging to distinct populations, there is an un-matching degree in LD correlation between these groups. In order to estimate the chance element, random re-sampling and permutation is carried out to draw the probability distribution, against which the real data would be tested.

Using Perl scripts running on a UNIX platform (see appendix 3, script 1); Initially all the pair-wise r^2 values are calculated for all markers, within each group separately, using the Expectation Maximization (EM) algorithm. Each group represents one of the populations that we want to establish the genetic distance between. Afterwards the Spearman's rank correlation coefficient (rho) is calculated for the r^2 values between the two groups. Each r^2 value in the first group is paired to the corresponding r^2 of the same marker pair in the other group. Spearman's rho estimates the association between paired samples and computes a test

of the value being zero. The measure of association has the range [-1, 1] with 0 indicating no association.

For estimating the probability distribution and P-values, a series of bootstrap sampling is carried out, each time constructing two new groups from the pooled sample of individuals from the two populations together. Individuals are randomly selected from the pooled sample, ignoring their ethnicity assignment, to create two random groups of the same sizes as the real groups.

For each of the two new random groups, pair-wise r^2 values are calculated, as well as the Spearman's rank correlation coefficient correlation, as done for the real groups. This process is repeated between (1000 to 50000 times).

The P-value for obtaining the result of the real data is calculated from the distribution of the permutations' rho values, as the number of rho values equal or less than the real data rho value divided by the total number of permutations.

Scatter plots of Spearman's rank correlation coefficients calculations, allele frequencies and LD statistics, for each pair of population comparisons, were done using the R statistics program (R Development Core Team 2005).

The bootstrap method: Bootstrap is a method to estimate various statistics and their reliability based on newly generated artificial samples. These new samples are obtained by drawings from the original sample. For a given sample of individuals new computer-generated samples may be formed by randomly selecting (with replacement) a desired number of individuals from the given set. Repeating this procedure an x number of times, provides a collection of x new samples which allows to estimate various indices of interest, their variations, and to apply subsequently inference statistical methods, e.g. significance tests or confidence interval estimation (Good 2006).

4.5. Results

After performing pedigree checks, there were 72 unrelated Masalit and 72 unrelated Hausa with data for 30 polymorphic loci, 64 unrelated Gambian individuals with data for 29 loci, 60 unrelated YRI individuals with data for 23 loci, and 60 CEU individuals with data for 24 loci. None of these loci showed any departure from Hardy-Weinberg Equilibrium.

4.5.1. Genetic Diversity within populations

Using the KOIND package (Kosman and Leonard 2007), several within-population diversity measures were calculated (Table 4.5.1). Values close to 0 indicate high uniformity, while large values indicate high diversity. 200 bootstrap samples were generated for each population's haplotypes. Measures of diversity were averaged over all bootstrap-derived estimates. High values of within-population diversity were found in all samples (Table 4.5.1).

Population\ Diversity index	Nei(Hs)	Muller(Mu)	Kosman expected(K)	Simpson(Si)
Hausa	-0.782	-0.789	-0.564	0.978
Masalit	-0.648	-0.655	-0.479	0.976
Gambians	-0.819	-0.827	-0.595	0.979
YRI	-0.744	-0.751	-0.538	0.977
CEU	-0.641	-0.647	-0.490	0.904

Table 4.5.1: Within- population diversity indices for all of the population groups analysed.

All studied groups displayed very high genetic diversity within themselves (Table 4.5.1). The same result was obtained by looking at haplotypes. Haplotypes were found to be highly diverse within groups, and there was a negligible proportion shared between the different

groups. This result is most likely due to the inadequacy of sample sizes for estimating genetic diversity. Previously, it was shown that the number of individuals to be used for estimating average heterozygosity should be large if a small number of loci are studied and the average heterozygosity is expected to be high. The number of individuals also needs to be large if the genetic distance of the two compared population groups is small (Nei 1978). This high within-population diversity makes it more difficult for clustering methods like *Structure* to correctly discern population groups.

4.5.2. Comparing allele frequencies between population pairs

To compare allele frequencies between pairs of population groups in the 5q31 region; first, minor allele frequencies in each group of unrelated individuals were calculated, then each marker was compared across a population pair using a 2x2 chi square test with one degree of freedom.

For pairs of African population groups, there were no significant differences in minor allele frequencies. The correlations in MAFs were found to be high between these population groups, with the Correlation Coefficient (R^2) ranging between 0.76 and 0.93 (Table 4.5.2a, Figure 4.5.2a).

Populations' pair compared.	*Correlation Coefficient (R^2).*	*Number of markers compared.*	*Number of individuals compared.*
Hausa vs Masalit	0.8263	30	72
Gambians vs Hausa	0.8792	29	64
Gambians vs Masalit	0.7548	29	64
YRI vs Gambians	0.87	23	60
YRI vs Hausa	0.9306	23	60
YRI vs Masalit	0.7756	23	60

Table 4.5.2a: correlation of Minor Allele Frequencies between pairs of African populations.

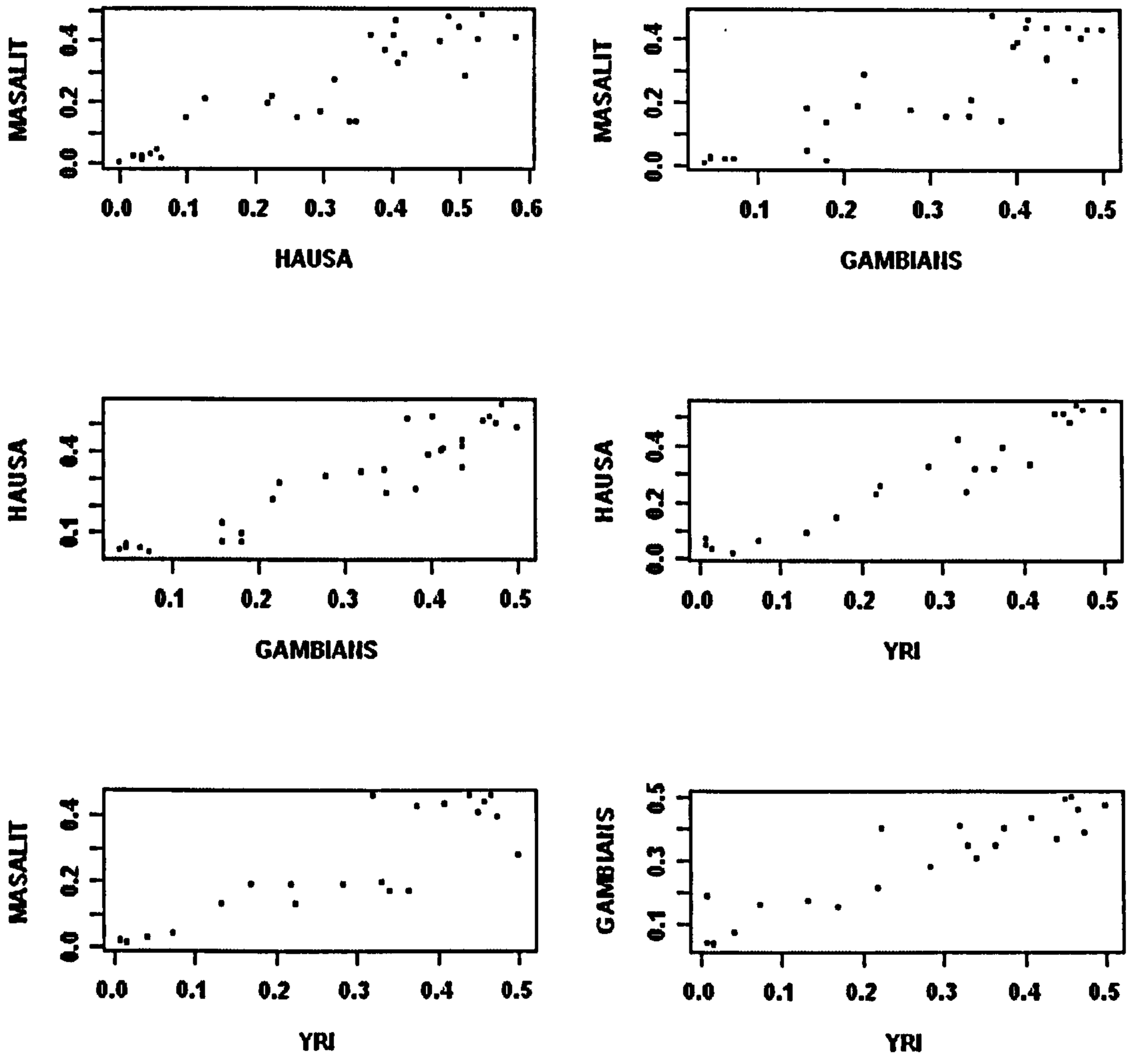


Figure 4.5.2a: The Correlation of Minor Allele Frequencies in comparisons between pairs of African population samples.

To explore whether the high correlations of allele frequencies between the pairs of African populations can be attributed exclusively to the choice of markers typed or not; publicly available data from the HapMap project for the same marker set in the CEU sample, was analysed and compared with each of the African population samples.

The high degree of correlation of minor allele frequencies observed between the African population samples was absent in comparisons involving HapMap CEU sample and employing a subset of the markers (Table 4.5.2b, Figure 4.5.2b), which suggests that the

similarities between African population groups result from the combined effect of close genetic relatedness and inadequacy of marker sets and sample sizes to unravel the differences in the genetic makeup of these groups.

Populations' pair compared.	*Correlation Coefficient (R^2).*	*Number of markers compared.*	*Number of individuals compared.*
CEU vs YRI	0.0066	23	60
CEU vs Gambians.	0.0217	24	60
CEU vs Hausa.	0.0143	24	60
CEU vs Masalit.	0.0755	24	60

Table 4.5.2b: Correlation of Minor Allele Frequencies between a population of a European origin (HapMap-CEU) and four African populations.

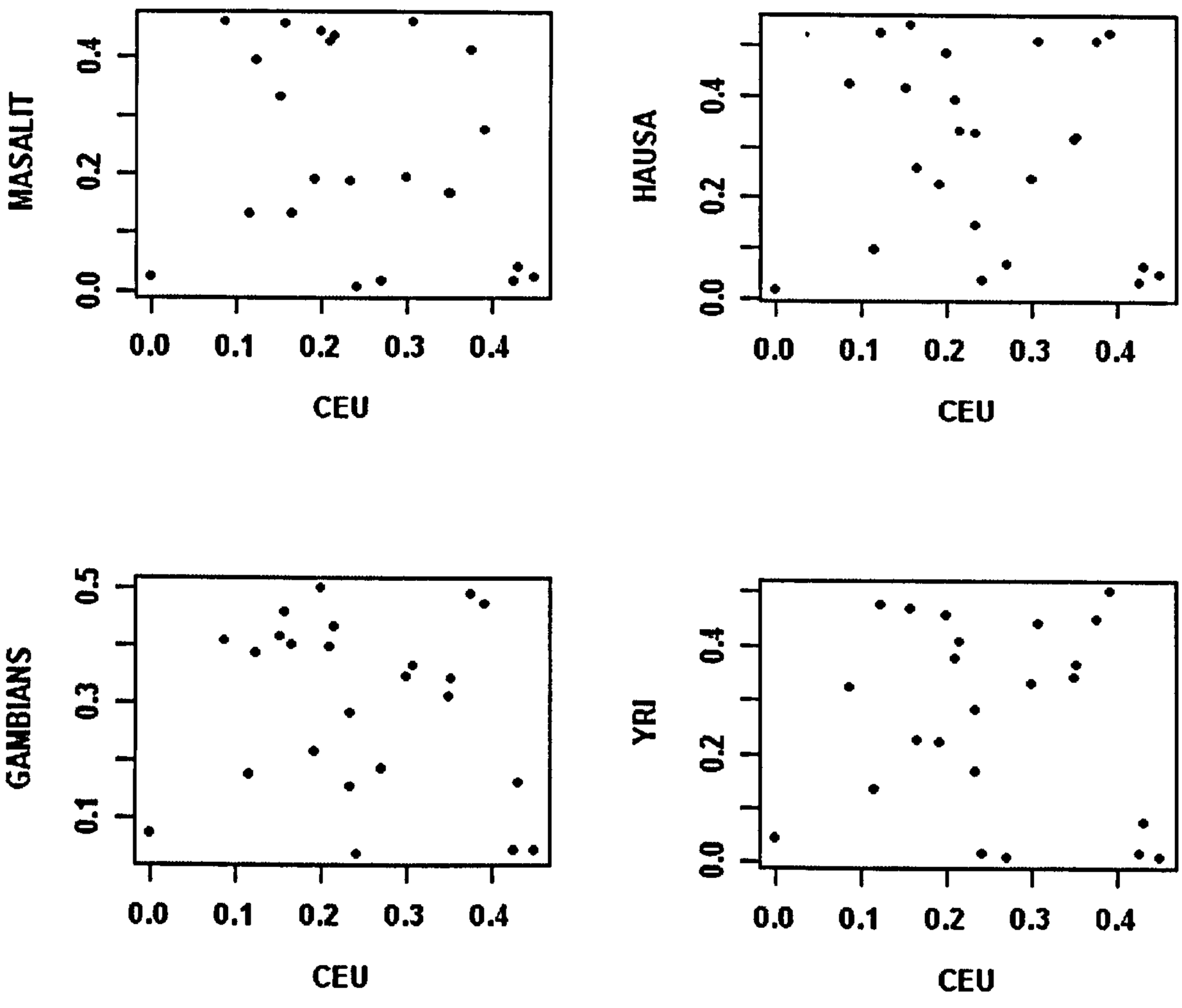


Figure 4.5.2b: comparing minor allele frequencies between the CEU sample and the four African population samples.

4.5.3. Comparing LD quantity and pattern between population groups

To explore whether the LD in the studied populations is quantitatively comparable, I describe some LD summary statistics. The average, variance, range and the median of pair-wise r^2 values over the whole set of loci in each population were calculated (Table 4.5.3a).

	LD average (variance)	LD median value	LD range
Hausa	0.05(0.01)	0.021	1.8E-05, 0.96
Masalit	0.05(0.01)	0.02	1.0E-05, 1
Gambians	0.04(0.01)	0.014	1.9E-17, 1
YRI	0.04(0.01)	0.014	1.2E-17 , 0.93
CEU	0.24(0.1)	0.108	9.2E-19 , 1

Table 4.5.3a: Summary of LD quantities in the five populations of the study.

LD was found to be quantitatively very similar between the African population groups. The median and average of LD values, as well as their variance and range is comparable for these population groups.

The CEU sample, as expected, harbors more LD than the African populations. It is evident from table 4.5.3a and figure 4.5.3b the higher LD values in CEU relative to the African population samples. The well known "out of Africa" bottleneck ≈200,000 years ago (Cann, Stoneking et al. 1987) reduced the genetic diversity of modern humans in Asia and Europe dramatically, leading to a higher LD in Europeans, as observed in several previous studies.

In contrast to the quantitative similarities in LD values between the African groups, patterns of LD appear to differ between these groups, as suggested from figure 4.5.3a.

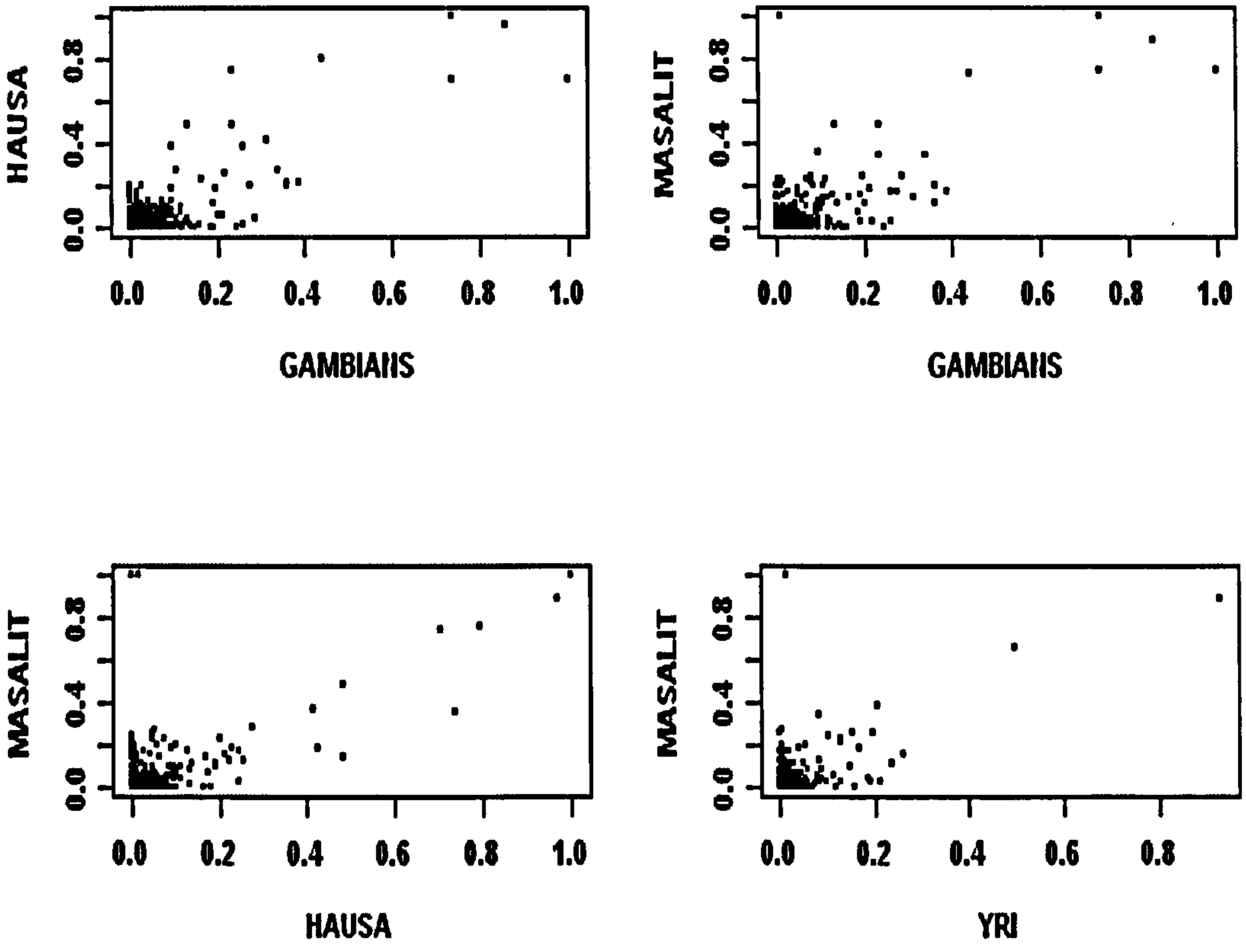


Figure 4.5.3a: Comparisons of r^2 values between pairs of African populations (each dot represents the r^2 value of a marker pair in one population and its corresponding value in the other population).

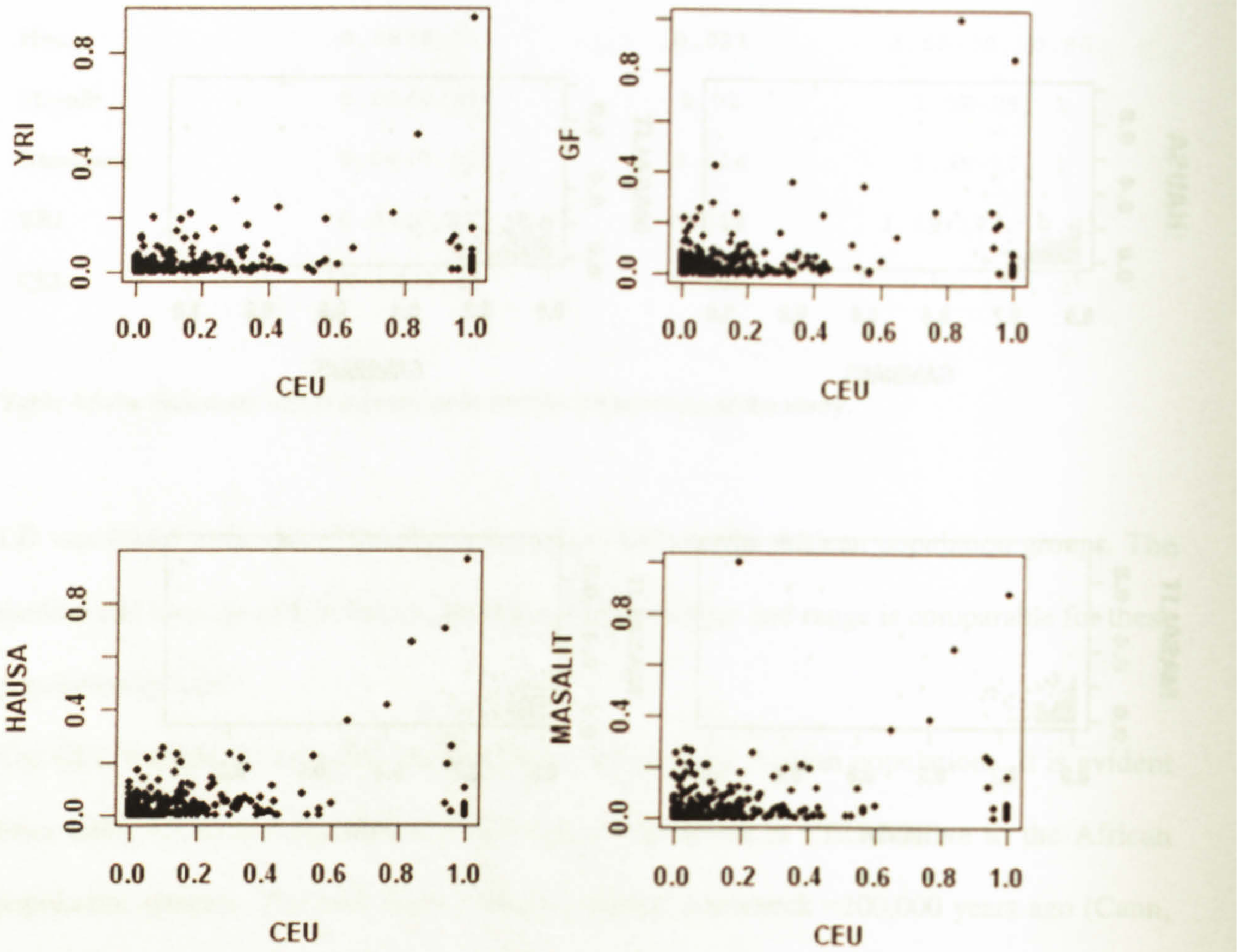


Figure 4.5.3b: Comparing r^2 values between CEU and African populations (each dot represents the r^2 value of a marker pair in one population and its corresponding value in the other population).

While no differences were observed between African population samples from comparisons of MAF and LD quantity, it became apparent there are some differences when examining their LD patterns. This observation merited further exploration to see if it is what chance alone would dictate -due to the inherent higher variability in LD statistics- or whether it is a true reflection of the genetic distances between these groups and could be usefully employed in capturing their genetic differentiation.

To address the above questions I used the following approach: first, to compare the LD between any two population groups for a particular set of markers, pair-wise r^2 values were calculated for all the marker pairs within each population group. The LD values in the two groups were then matched and their correlation was calculated using Spearman's rank correlation coefficient (rho), which is a non-parametric measure of correlation in which raw scores (pair-wise r^2 values) are converted to ranks within each sample, and the differences between the ranks of paired observations in the two samples are calculated.

To quantify the chance element in the observed correlation value, the null hypothesis had to be tested. The null hypothesis states that the observed low correlation in LD values between the two population groups is due to normal sampling fluctuation because of the inherently high variability of the LD statistic, and not due to the different ancestry of the groups which means than the two population samples came from the same pool of individuals and are actually no more different than any other two samples drawn from the combined sample. To test whether an observed value of rho is significant is to calculate the probability of it being greater than or equal to the observed rho given the null hypothesis. In order to achieve this, I chose to use a permutation test because it is generally considered superior to traditional methods of calculating significance. The individuals in the two compared samples were pooled together and two new bootstrap samples were randomly chosen from this combined sample, effectively switching group membership for some individuals. r^2 and rho were then calculated for the new random samples as described above. The bootstrap re-sampling was then repeated a great number of times to get the probability distribution of the rho values, from which the real data P value was calculated.

Analyzing the Hausa and Masalit; the Spearman's rank correlation coefficient (rho) of r^2 values in the two groups was found to be (0.411878). When a 10,000 permutations were carried out, correlation coefficients from these permutations, had a normal shaped distribution (Figure 4.5.3c), which supports the assumption of randomness and justifies the way I chose for calculating the P value. From this distribution the P-value of observing the real-data rho value was found to be 0.016. That is to say, out of the 10,000 permutations carried out, 160 had a correlation value that was equal to or less than the real data (represented in figure 4.5.3c by the left tale of the distribution from the red arrow).

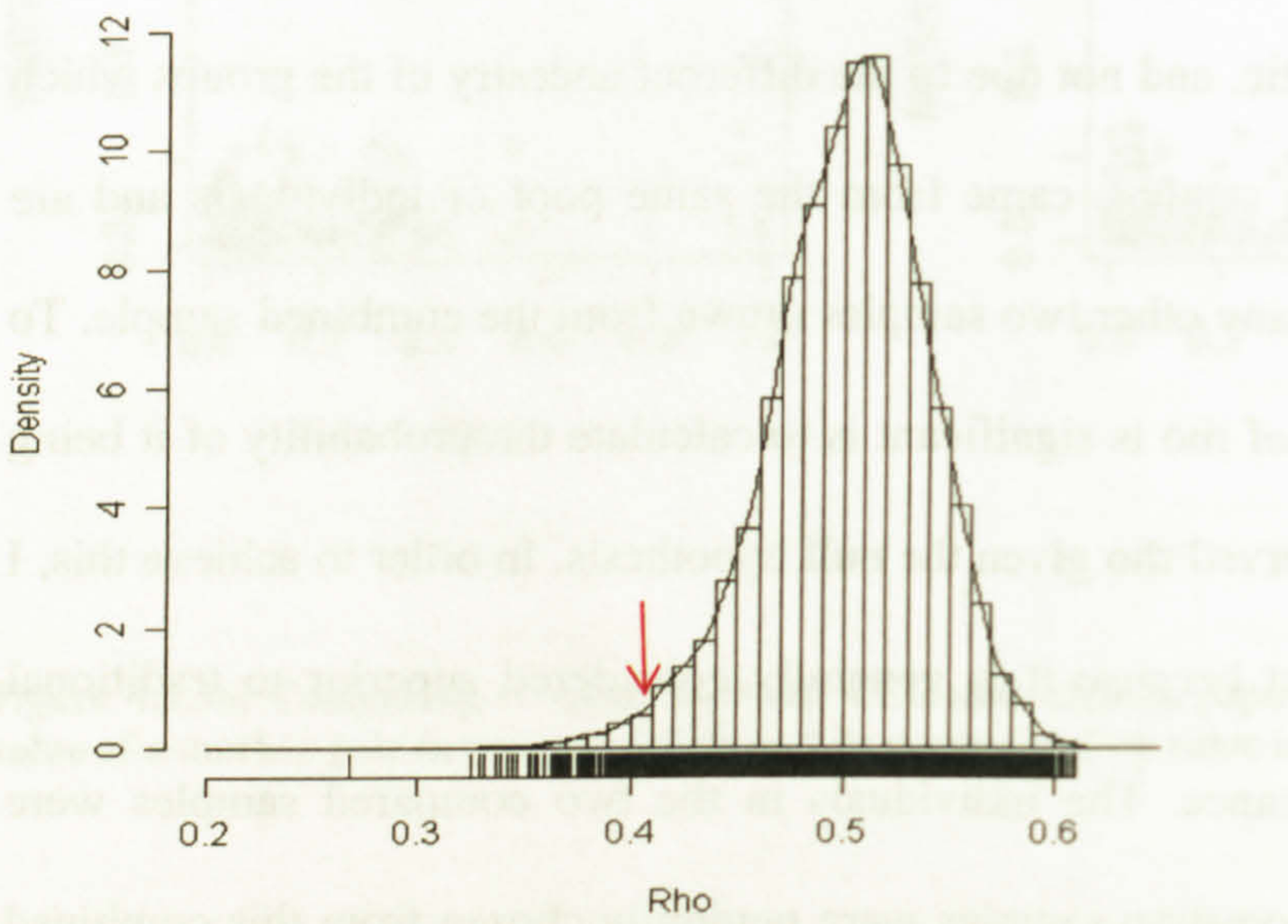


Figure 4.5.3c: Distribution of the Spearman Correlation Coefficient (rho) values for 10000 permutations of the Hausa and Masalit LD-pattern comparison. Red arrow indicates the correlation value of the real data.

Number of permutations	*Hausa/Masalit P-values*	*YRI/CEU P-values*
50000	---	0.00004
10000	0.016	0.0001
8000	0.014	0.00013
6000	0.013	0.00017
4000	0.013	0.00025
3656	0.016	---
3000	0.017	0.0003
2000	0.013	0.0005
1000	0.011	0.001
100	0.02	0.01

Table 4.5.3b: P-values obtained from different number of permutations of the LD-based genetic distance analysis.

I ran the analysis several times with a different number of permutations each time, to determine the minimum acceptable number. There were two criteria that I sought to fulfil. First, the number of permutations had to be large enough to generate at least one instance of rho that is equal to or less than the rho value of the real data. Second, the minimum acceptable number of permutations should ideally be at the point where the P-value starts to level off with no significant decrease in the P-value with increasing the number of permutations.

When carrying out the comparison between the Hausa and Masalit samples using different numbers of permutations to estimate the P values, 1000 permutations were enough to generate several random samples with less correlation values than the real data (less than 0.41). Permutations above that did not much increase the accuracy of estimating the P value (Table 4.5.3b). On the other hand, when the CEU and YRI were analysed, the P-value

decreased steadily with increasing the number of permutations up to 10,000 (Table 4.5.3b). Even this number of permutations was not enough to generate a single rho value of less than 0.16 which is the correlation value between the real groups. 50,000 permutations were required before lower-than-real-data rho values were obtained by chance.

From the above, it appears that the number of required permutations differs with the data set specifics like the number of markers, the sample size and divergence between populations from which samples are taken. Therefore, it might be reasonable to attempt several levels of permutations before deciding on the P value.

The LD pattern difference between the Hausa and Masalit appeared to be greater than what chance would dictate. To further test this, I used the genotypic data from two additional African populations, the Gambian population sample comprised of 64 unrelated individuals (Luoni, Forton et al. 2005), and the HapMap data for 60 unrelated Yoruba of Ibadan, Nigeria (YRI).

Populations' pair compared.	*Spearman Correlation Coefficient (rho).*	*P-value.*	*Number of markers compared.*	*Number of individuals compared.*
Hausa vs Masalit	0.411878	**0.015698**	30	72
Gambians vs Hausa	0.275096	**0.035796**	29	64
Gambians vs Masalit	0.313772	0.076692	29	64
YRI vs Gambians	0.196855	**0.031097**	23	60
YRI vs Hausa	0.087956	**0.0008**	23	60
YRI vs Masalit	0.2897	0.435356	23	60

Table 4.5.3c: LD-based genetic distance estimation between pairs of populations of African origin. Each analysis is of 10,000 bootstrap permutations. P-values < 0.05 are shown in bold.

The number of markers used in each population pair comparison, represents the intersect of markers typed in both populations. The number of individuals was that of the lesser group. In total six comparisons were carried out with all the pairs of African populations. Out of these, four comparisons yielded significant results, when significance level was set to 0.05 (Table 4.5.3c).

I explored whether it is reasonable to compare results from different analyses. Figure 4.5.3d displays the distribution of rho values for all population pairs analysed, as box plots next to each other. Overlap in the distributions indicates the possibility of comparing results from different analyses.

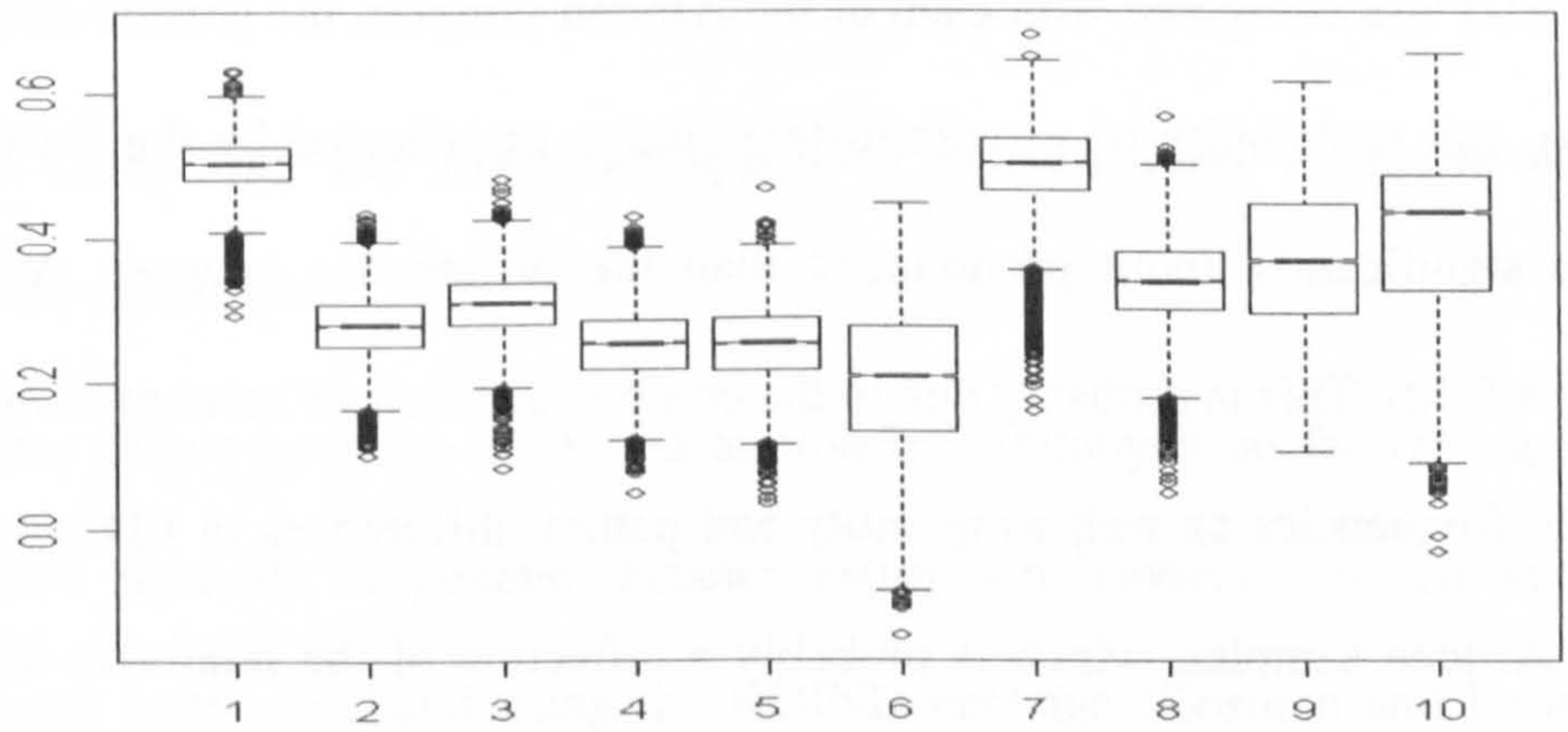


Figure 4.5.3d: Comparing the distributions of rho values for all population-pair comparisons. On the x axis the between-populations comparisons are ordered as follows (1-Hausa&Masalit. 2-Gambians&Hausa. 3-Gambians&Masalit. 4-YRI&Gambians. 5-YRI&Hausa. 6-YRI&Masalit. 7-YRI&CEU. 8-CEU&Gambians. 9-CEU&Hausa. 10-CEU&Masalit.). On the Y axis the rho values are displayed.

All comparisons between African populations showed the two compared groups to be significantly different from each other at the 0.05 significance level, except when comparing the Masalit with the Gambians and with the Yoruba (Table 4.5.3c). Higher sample diversity in the Masalit cannot explain this result, as the within-population diversity in the Masalit is

comparable to that of the other samples. It is more likely attributed to low resolution of this marker set in the Masalit. The least similar groups were those of the Hausa and YRI (rho = 0.088, P-value = 0.0008). This apparently larger genetic distance relative to the other population pairs is interesting. The fact that Hausa being part of a long history of Nigerian populations, and possibly admixing with populations in the South like YRI, might lead to the assumption that the relationship between the Hausa and YRI might be a closer one than that of the other population pairs. This shows that, while comparing LD patterns might have higher sensitivity in discerning genetic distance, it is not perfectly correlated with it and it is not the final answer in determining the between-populations genetic distances, rather it should be taken within the context of other evidence.

When a European population sample (Utah residents with ancestry from Northern and Western Europe-CEU) was compared with each of the African samples, the genetic distance as reflected by rho, and probability of genetic differentiation as reflected by the P-values; were found to be significantly more pronounced than the differences between African populations (Table 4.5.3d). This is probably due to the combined effects of more pronounced differences in allele frequencies as well as quantity and pattern differences in LD between the European and African samples, which is probably a reflection of the relatively distant ancestry between CEU and the African groups.

Populations' pair compared.	*Spearman Correlation Coefficient (rho).*	*P-value.*	*markers compared.*	*individuals compared.*
CEU vs YRI	0.160754	0.00004	23	60
CEU vs Gambians.	0.060142	0.00002	24	60
CEU vs Hausa.	0.201406	0.00044	24	60
CEU vs Masalit.	0.063556	0.00012	24	60

Table 4.5.3d: Genetic distances estimated with the LD-based method between HapMap-CEU and several African populations. 50,000 permutations were carried out per analysis.

4.5.4. Comparing the LD-based approach against some available metrics of genetic distance estimation

In order to see how the LD-based approach I employed so far fairs against some of the available methods of genetic distance estimation, several between-populations diversity measures were calculated using the KOIND package (Kosman and Leonard 2007) (Table 4.5.3).

Even though the KOIND package is intended for the haploid genomes of plant pathogens, I found it extremely useful for the intended LD-based analysis. It suited the nature of the datasets I am using, as it is designed for data with strong LD between its markers.

	The Nei distance(N)	The Nei coefficient of differentiation(Gst)	The Kosman-Leonard expected distance(KL)	The Rogers distance(R)
Hausa vs Masalit	0.006	-0.009	0.078	1.000
Gambians vs Hausa	0.004	-0.006	0.063	1.000
Gambians vs Masalit	0.005	-0.011	0.083	1.000
YRI vs Gambians	0.003	-0.008	0.059	1.000
YRI vs Hausa	0.002	-0.005	0.050	1.000
YRI vs Masalit	0.005	-0.008	0.073	0.999
CEU vs YRI	0.031	-0.045	0.182	1.000
CEU vs Gambians	0.026	-0.037	0.180	1.000
CEU vs Hausa	0.030	-0.040	0.185	1.000
CEU vs. Masalit	0.034	-0.047	0.197	1.000

Table 4.5.4: Between-populations diversity indices for each population pair analysed.

The first three columns in table 4.5.4 give results of metrics that depend on allele frequency comparisons. Their results show larger genetic distances between CEU and the African population groups than between African groups. This trend agreed with what was shown by the LD pattern analysis in the previous section.

On the other hand, Rogers distance (R) which is an approach that utilizes the full haplotypic information of all typed markers (the last column in table 4.5.4), resulted in maximum or near maximum values across all comparisons with no difference in the estimated genetic distances across or within continents. This probably represents an overestimation of the between-populations genetic distances and might indicate the unsuitability of this kind of metric to analyse these data sets.

The program **STRUCTURE** 2.0 (Pritchard, Stephens et al. 2000; Falush, Stephens et al. 2003) is one of the most widely used software packages for determining population stratification. When I used it to identify stratification in the combined sample which is comprised of the two Sudanese populations, the Gambian sample and the HapMap YRI and CEU, there was no obvious distinction between the populations in spite of their well known different ancestry (Figure 4.5.4a). Although the CEU were better identified as a cluster compared with the African populations, still there was a significant degree of wrong assignment of the individuals in the sample. This observation suggests that the number of genotyped markers is inadequate for **STRUCTURE** to identify population divergence between these groups, the effect of inadequate markers becomes more pronounced, the more closely related the populations are to each other, as is observed between the African groups.

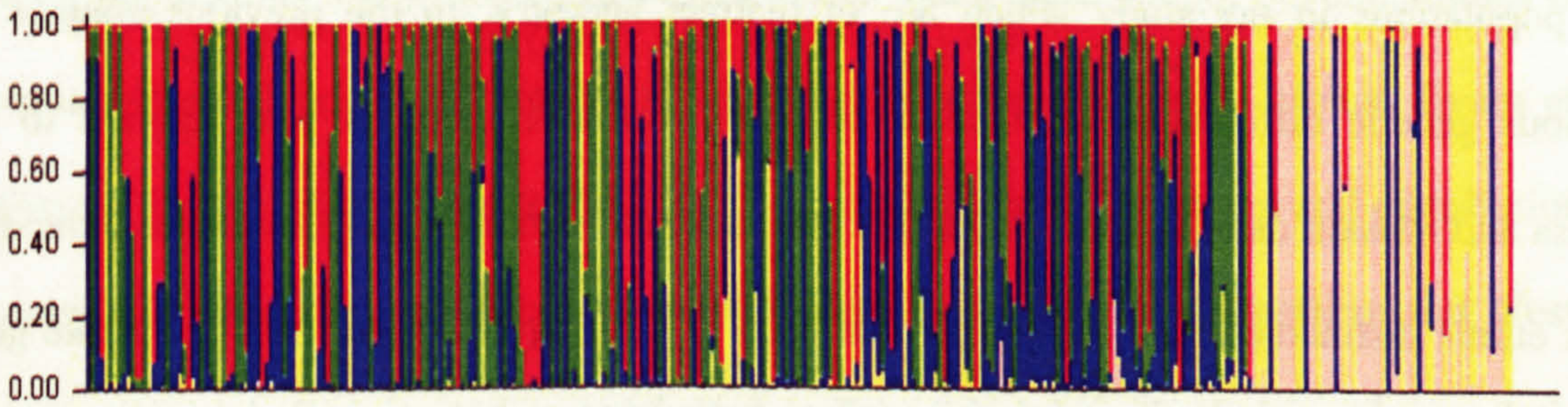


Figure 4.5.4a: STRUCTURE Bar plot of individuals' ancestry under no admixture model and assuming five populations of origin of the combined unrelated Hausa, Masalit, Gambians and HapMap YRI and CEU samples with data for 30 markers in the 5q31 region. In figure each vertical bar represents an individual. Population groups were ordered as Hausa, Masalit, Gambians, YRI, and CEU on the x axis. On the y axis the proportions of the individuals ancestry assigned to the five populations are shown with different colours.

When the markers analysed were increased in number to 80 in total by using genotypes from several other genomic regions (see Chapter5), **STRUCTURE** still failed to distinguish between the African population groups (Figure 4.5.4b).

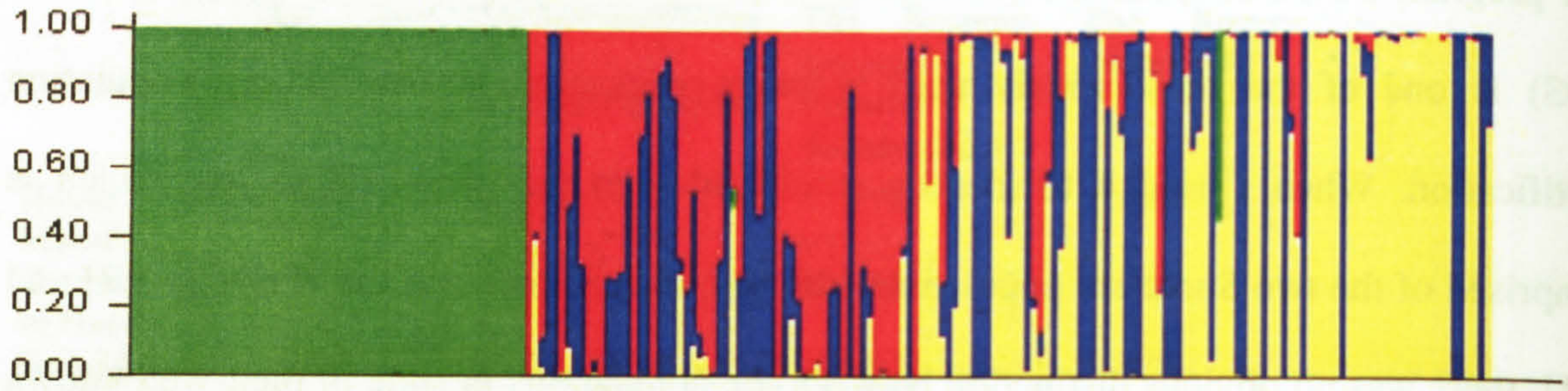


Figure 4.5.4b: STRUCTURE Bar plot of individuals' ancestry under no admixture model and assuming four populations of origin of the combined unrelated Hausa, Masalit, and HapMap YRI and CEU samples with data for 80 markers. In figure population groups were ordered as CEU, YRI, Hausa and then Masalit. On the x axis each vertical bar represents an individual. On the y axis the proportions of the individuals ancestry assigned to the four populations are shown with different colours.

4.6. Discussion

In this chapter I have tried to reveal, and possibly quantify the genetic differences between the populations in my study which are of distinct ancestry. In the previous chapter no obvious genetic differences were observed when available methods were applied to the Hausa and Masalit data, in spite of the fact that there are several lines of evidence indicating their ethnic distinctness and suggesting that each of these groups represents a separate gene pool. The case for Hausa and Masalit ethnic differences draws substantiation from several sources. In spite of their geographic contiguity – they inhabit neighbouring villages in a remote area in Eastern Sudan- they have diverse historical accounts of their origins with a well known recent history. Their languages belong to different linguistic families. They represent relatively isolated populations with a limited contemporary gene flow and no admixture with other ethnic groups. In each village the founders were all derived from the same original gene pool, and the pedigree structure showed that within each village, there was a high degree of relatedness between individuals from different families. The whole village can be divided into few clusters of families with close kinship ties between their

members. Therefore, founder effects and genetic drift are expected to have played a big part in the genetic differentiation between these populations. In a previous study conducted in the same area of Eastern Sudan, analysis of two nearby villages occupied by the related ethnically uniform Masalit group, found different susceptibility loci in the two villages to the same phenotype. This result was attributed to chance founder events that carried specific susceptibility alleles into each village (Miller, Fadl et al. 2007).

The genetic similarities between the Hausa and Masalit as well as the other African groups were reflected in the very high correlations of allele frequencies across African populations (R^2 between 0.8 and 0.9) (Table 4.5.2a). And although the quantity and extent of LD across all African populations were found to be comparable as well (Table 4.5.3a), when the pair-wise r^2 values were compared for each pair of population groups, the qualitative difference in LD pattern between different groups was obvious. Scatter plots comparing r^2 statistics between populations are presented in Figure 4.5.3a and Figure 4.5.3b. Several studies have previously highlighted the differences in LD patterns across population groups. Evans et al. (Evans and Cardon 2005) have found considerable variation in the extent and distribution of pair-wise LD when they compared samples of East Asians, African American and Western European descent. Recent studies conducted in four non-urban Sardinian sub-populations, indicate that, even neighbouring villages of sub-isolates derived from the same founding genetic pool, may have contrasting extents and patterns of LD (Angius, Bebbere et al. 2002).

The observation of different LD patterns between the populations of my study raised the question of whether this apparent difference in LD pattern is due to chance fluctuation of random sampling of an inherently variable statistic (LD statistic sampling is potentially more variable than that of allele frequency), or if it is a true mark of the different ancestry of the samples. The relatively low correlation in LD patterns also raised the question of whether this property is helpful in determining the genetic distance between population groups and if

it can be employed in some sort of metric to calculate it. Although the notion of contrasting LD patterns, has been introduced before to compare case and control samples, for the purpose of mapping and identifying disease susceptibility loci (Zaykin, Meng et al. 2006); it has not been used before to discern genetic differentiation between populations of diverse ancestries, with employment of the permutation approach, to take the analysis a step beyond the graphical representation, and allow for statistical quantification of the differences between groups compared.

To address the above questions, I performed LD pattern comparisons using Spearman's rank correlation coefficient (rho) and a bootstrap permutation approach to determine significance. The results agreed with the general principle of larger genetic distance between populations from different continents than those within the same continent. As expected the degree of correlation of within-African-populations r^2 statistic was found to be higher than correlation statistics between CEU and African groups (Table 4.5.4c and Table 4.5.4d). Nevertheless, there still were significant differences between pairs of African populations. In a study of genetic variation among world populations; African populations were found to be more diverse than other continental groups and the largest genetic distance was seen between them and non-African populations (Watkins, Rogers et al. 2003). Nei et al. (Nei 1982) studied the genetic relationships of various races in each group of Europeans, Africans, and Asians, and found all European populations to be genetically close to one another, whereas many African tribes show large extents of genetic differentiation.

Four out of the six comparisons made between African populations yielded significant results. The groups that had the biggest difference were the Hausa and YRI, with a P value of 0.0008. The most similar groups were the Masalit and YRI, with a P value of 0.44. These quantitative estimates of genetic distances did not agree with what was expected from the histories of these populations. The Hausa being originally from West Africa is expected to

be more similar to the West African YRI than the Masalit population which has an East African origin. Therefore, although the LD pattern comparison was successful in highlighting genetic differentiation between groups and estimating the confidence in the results, whether the resulting P values express the degree of that difference remains an area for further work and discussion.

The extent of the genomic region used in the current analysis – a 650kb segment of the 5q31 region- lends itself to the proposed approach of exploiting LD information to discern genetic differences between populations. In most human populations, LD extends for relatively short distances, on the order of 10s to 100s of kb in most genomic regions (Reich, Cargill et al. 2001), but in some instances LD may extend to longer distances. Luoni et al. (Luoni, Forton et al. 2005) found that high LD tends to be more dispersed in African populations, as opposed to the close range at which high LD is observed in a European population. Over the same genomic region of my current analysis, they have found more examples of long-range LD in the Gambian population, where instances of high LD between SNP pairs were significantly more likely to span a distance of >200 kb. Whether that is due to positive selection or population demography like small effective population size or recent admixture, it can be utilized in distinguishing population groups from each other. Also it has been shown that useful LD extends over large genetic distances in isolated populations (Angius, Hyland et al. 2008). Although it was previously observed that global LD profiles of human populations show overall similarities corresponding to shared recombination patterns (Service, DeYoung et al. 2006); at this finer scale, there exist differences in LD patterns that might reflect variations in demographic histories between populations.

I considered, but eventually excluded, the decline of LD with distance as a way of comparing LD pattern between groups, because in isolated populations LD could be present

between linked and unlinked markers (Abecasis, Ghosh et al. 2005), suggesting that there is more to the pattern than the decay in LD, that could be affected by population structure.

The reason why I chose to use r^2 instead of other measures of LD like absolute D', was because r^2 reflected marker allele frequencies more closely than the other measures, thus combining more of the information contained in allele frequency differences with that from the LD between markers. It is also the standard measure of whether the LD between two markers is sufficient for detecting phenotype associations. In a recent study, D' displayed a ceiling effect with most points reaching a maximum value of 1.0 in one of the compared populations. This resulted in high variability for this measure even when groups of similar ancestry were compared. This suggested that the observed lack of concordance between groups was an effect of the measure used rather than different group ancestries (Evans and Cardon 2005). Therefore, the authors suggested that the r^2 measure might be more useful in trans-population disease gene mapping than the pair-wise D'.

An interesting aspect to this LD-based approach is the possibility of using LD relationships between pairs of markers as an alternative to identifying populations-differentiating marker sets, when none of the markers typed is significantly different in frequency across compared groups. LD values with the highest disparities between groups could be used for this purpose, especially if these disparities consistently hold for one population when compared with others. An example of that is the relationship between marker rs1295686 and rs200541. These two markers have a comparable Minor Allele Frequencies in all the population groups that were looked at. Consequently neither of these SNPs could be used as a population-differentiation marker on its own. On the other hand, when the LD relationship between these two markers was explored, they were found to be in perfect LD in the CEU population,

which was in stark contrast to their very low LD value in all other populations considered (Table 4.6).

	MAF(rs1295686)	*MAF(rs200541)*	r^2
CEU	0.23	0.23	**1.00**
Masalit	0.2	0.18	0.05
Gambians	0.29	0.17	0.08
Hausa	0.3	0.15	0.04
YRI	0.28	0.17	0.08

Table 4.6: Two markers with no significant difference in their Minor Allele Frequency (MAF) in different populations, exhibit high disparity of their LD values between CEU and other populations.

I applied some of the available genetic distance estimation metrics to the data, in order to compare them with the results of the LD-based analysis. Measures of diversity within and distance between populations based on frequencies of individuals' haplotypes, like Rogers distance(R) (Kosman and Leonard 2007), overestimated actual diversity within and differences between populations (Table 4.5.1, Table 4.5.4). This bias can be explained as follows. The number of individuals tested, is limited and considerably less than the number of theoretically possible haplotypes (2^x in the case of x independent binary characters) even for relatively small values of x. Therefore, it is very likely that nearly all sampled individuals are of different genotypes. Actual population diversity might be much lower because of possible similarity between overall allelic patterns of individuals not sampled. Similarly, the Rogers distance may overestimate actual difference between two populations because of the high probability that none or a very small number of individuals in the two compared samples will have identical genotypes. Methods based on comparing haplotypes within and

between populations can also be hampered by the degree of uncertainty in haplotype phase inference, especially when extending over a 650 kb region.

Given enough time following populations' division, isolated groups become genetically more divergent with time, either by acquiring new mutations, or when the original haplotypic backgrounds on which existing variants lie change in abundance by drifting upward or downward in the population. That plus the reshuffling caused by recombination, affect the associations between these variants in ways that are specific to each population group. These changes can either take place over long time periods due to genetic drift or can happen over a relatively short time span when local environmental factors exert selective pressure on a functional variant sweeping nearby neutral polymorphisms up with it. The effects of all of these factors could potentially be reflected in allele frequencies profiles, but I found the effects on allele frequencies to be obscured in my datasets, probably due to the fact that most SNP data have been obtained by choosing high frequency markers from publicly available databases. This ascertainment bias complicates any downstream analyses based on allele frequency differences. However, these processes by which SNPs have been selected would probably bias allele frequencies more so than levels of LD observed in the data. While high frequency variants are more likely to be old and shared between population groups, consequently displaying little frequency differences between compared groups; these high frequency variants are more valuable in highlighting historical recombination events because of their higher resolution. This might suggests a potentially higher sensitivity of an LD-based approach in determining genetic distances than one that is based on allele frequency alone.

The effects of the ascertainment bias were clearly manifest in the failure of most available methods that rely on differences in allele frequencies to highlight any significant genetic differentiation between studied population groups (Table 4.5.3). The number and characteristics of markers typed might not have had enough resolution to distinguish these populations from each other. This suggested an inadequacy of the combination of method and data for unravelling genetic distinction between the population samples.

Clustering algorithms like those employed in Arlequin and Phylip gene tree construction did not manage to correctly assign individuals to proper clusters (Chapter 3). ***STRUCTURE*** failed to discriminate between all populations using data from 30 markers, and even when markers were increased to 80, the program only managed to distinguish the CEU population, leaving the African populations indistinguishable from each other.

Bamshad et al. (Bamshad, Wooding et al. 2003) indicated that a small sample of loci doesn't typically provide sufficient power to detect population structure. The low resolution of allele frequency information content of loci used in the analysis was not sufficient to discern populations' genetic differences, especially when considering that linked markers give nonindepentant information and are therefore less informative than are the same number of unlinked markers. Rosenberg et al. (Rosenberg, Burke et al. 2001; Rosenberg, Pritchard et al. 2002) tested the effect of the number of loci on ***STRUCTURE*** clustering results and found that accurate clustering of individuals from extremely closely related populations can only be achieved when large numbers of markers are used.

Furthermore, under the linkage model, ***structure*** inferred that all individuals from both groups are admixed. ***STRUCTURE*** authors observed from simulations based on a variety of demographic scenarios, that this kind of result indicates background LD in the data

complicating the analysis. They suggest that genuine admixture should be asymmetrical, affecting one population more than the other (Falush, Stephens et al. 2003). ***STRUCTURE*** accounts for the correlations between linked markers by assuming it is due to admixture, but does not implement a way to deal with background LD (that is LD generated by genetic drift and expected to be strong between syntenous loci separated by few cM). Consequently, for ***STRUCTURE*** to make meaningful inferences, it can only deal with data from unlinked or weakly linked loci. Kaeuffer et al. (Kaeuffer, Reale et al. 2007) showed that *structure* could be sensitive to even a rare pair of loci in strong LD, resulting in clustering bias and the generation of spurious results. Pritchard and Wen (2004) advised users against using loci separated by less than 1 cM.

To summarize the above, commonly used measures of population diversity or genetic distance consider either allele frequencies or haplotype frequencies. The allele frequency based methods, require large numbers of markers to be typed at unlinked loci, while the haplotypes based methods require a large number of sampled individuals from each group, to accurately estimate diversity. So using methods based on allele frequency comparison may not be the most efficient approach in this setting. Not only does it not utilize the full information content of the data, but some methods recommend the exclusion of pairs of strongly linked loci that potentially bias the results.

The results of LD-pattern comparison supported oral traditions and historical accounts of the diverse origins of the populations of my study. And although it is a problem specific solution that I had to come up with in order to analyse the data available, it has the potential, after further testing and method development, of being applied and generalized to make a sensible testing framework relevant in association studies. The apparent relative sensitivity of this test to detect genetic distance did not safeguard against missing some cases, when either, the

genetic distance was too small, or the dataset was inadequate to discern genetic differentiation.

In order to develop the LD pattern comparison approach into a fully fledged applicable method of genetic distance estimation, rigorous testing for robustness and false discovery rate estimation needs to be carried out which might be more suitable for further future work beyond the scope of this thesis. However, as it stands it is an interesting observation that raises several questions about a number of situations and areas of opportunity where the application of a method based on this concept might be helpful and relevant to substantiate conclusions about genetic differentiation. For example when collecting from nearby villages inhabited by the same ethnic groups it would be helpful to make a judgement about whether they can be treated as the same sample (Miller, Fadl et al. 2007), especially when local environmental pressures might have accelerated their differentiation. This can be done with a modest data set, without the need for a large number of unlinked markers to be typed. In situations where limitations are created by amount and type of data, i.e. a few tightly linked markers in a small genomic region, typed in a few individuals, without the advantage of genome wide data, it is likely that no single best method can be recommended for the estimation of genetic differentiation, and several analyses could be considered, in conjunction, to form a judgement in each specific case.

4.7. Conclusion

The idea of comparing LD patterns between groups of populations, combined with a permutation approach, proved successful in genetically distinguishing the Hausa and Masalit. Other African and non-African populations were included to further test the

approach. Four out of the six comparisons between African populations were significant, and the largest differentiation was found between populations across continents.

I have shown that the genetic differentiation between populations can be highlighted using information contained in the LD patterns, when the data set is limited, rendering available methods not sensitive enough to infer genetic divergence. I managed, to a large extent, to tease out the distance between populations as predicted by their self specified ethnicities.

Although this approach shows promise as a useful addition to existing methods of estimating genetic distance when analyzing similar datasets, and it is of special relevance in the context of association studies; yet the full development and rigorous testing of such a method is beyond the scope of this thesis. I've decided to go back to the empirical data to explore, in more depth, the question of characterizing the signals of positive selection in the Hausa and Masalit by genotyping the HBB genomic region.

Chapter 5:

Genetic polymorphism and positive selection patterns in the β-globin region in the Hausa and Masalit of eastern Sudan.

5.1. Abstract

Identifying signals of positive selection in a genomic region of interest may offer important clues in the search for disease modifying variants. In previous work I found no clear signals of selection in the 5q31 genomic region in the Hausa and Masalit, so in this chapter I studied the β-globin region, as a classic example of a locus under positive selection, to provide a bench mark for analysis of other regions of the genome.

I genotyped 26 markers, including the HbS polymorphism and six classical RFLP markers, in 48 unrelated Masalit individuals and 47 unrelated Hausa individuals. A subset of the samples, those found to be carrying the sickle allele (12 Masalit and 9 Hausa), were further genotyped for another 37 markers spaced across a genomic area measuring 2Mb around the HbS polymorphism. I characterized the β-globin region in terms of haplotype structure, LD, genetic diversity, and signals of positive selection, with special focus on studying the HbS polymorphism, its frequency and haplotypes.

The Hausa group displayed a very distinct selection signal in the β-globin region. However, the detection of this signal was conditional on including the functional variant for sickle haemoglobin in the analysis. The observation that sickle haemoglobin haplotypes, had much higher frequency than other haplotypes across the region, prompted an in-depth look at how far these haplotypes extend, and introduced the possibility of utilizing such phenomenon in

detecting positive selection in the genome, especially when the functional variants might be missed out in the genotyping efforts.

5.2. Objectives

- Characterize the β-globin region in chromosome 11 in terms of haplotype structure, LD, and genetic diversity, in the Hausa and Masalit of eastern Sudan. Compare it to data from other populations.
- Study HbS variant, its frequency and haplotypes and attempt to link that with the classical HbS haplotypes, and what is known about malaria epidemiology in the two populations.
- Characterize the signal of positive selection in the region, and explore the extent of its effects on the patterns of genetic diversity and haplotype structure.
- Further explore the effects of demography on genetic variation in the region, and use insights gained from the β-globin region to shed light on genetic variation in the 5q31 region.

5.3. Introduction

Regions of the human genome containing disease resistance genes may be under considerable selective pressure if the disease phenotype leads to a reduction in fitness. Even very small fitness effects may, on an evolutionary time scale, leave a very strong pattern. Therefore, in theory it may be possible to identify putative genetic disease factors by identifying regions of the human genome that are currently under selection. Inferences regarding selection have therefore been used extensively to identify functional regions or

protein residues (Blanchette and Tompa 2002).(Sawyer, Wu et al. 2005). One practical and immediately applicable benefit of looking for signals of positive selection in the genome is in reducing both the economical and statistical costs involved in marker choice for association studies.

Challenges to interpretation of the genetic variation patterns in the Hausa and Masalit

In chapter 3, the interpretation of the positive selection patterns in both Hausa and Masalit genetic data proved to be problematic due to the presence of too many variables beside the putative disease modifying variant(s) under investigation in the 5q31 region.

This complexity stems from the relative contribution of any of the following factors to the snapshot of observed genetic variation data in these populations:

- Migration, founder effects and genetic drift.
- Population growth rate.
- Population substructure, inbreeding and admixture.
- Natural positive selection.
- Choice and coverage of markers.

One way of reducing the uncertainty is controlling for some of the variables by comparisons with other genomic areas where one or more of these factors are well defined. Uncertainty about whether or not there is a selective pressure shaping the patterns of the polymorphism data, could be addressed by studying the β-globin region on chromosome 11, where the genetic polymorphism under selection is well studied and described. The sickle cell allele in the β-globin region is the primary example of a polymorphism driven up by the selective pressure of *P. falciparum* malaria. This approach makes available prior knowledge of the

functional variant and its attributes, i.e.: position, frequency, LD and haplotypic relationship with other markers in the region.

A clearer picture of positive selection effects on the genetic variation in the Hausa and Masalit should emerge from studying the β-globin region. Applying knowledge gained from, and looking for patterns recognized in the β-globin to the 5q31 region, might help to better interpret its genetic variation results.

How positive selection affects the genetic variation in a genomic region

When a new beneficial mutation increases in frequency in a population because of natural selection, the genetic variation in neighbouring regions will be affected. The level of variability will be reduced, the level of LD increased, and the pattern of allele frequencies will be skewed (Braverman, Hudson et al. 1995).

The HbS polymorphism and Malaria

Haemoglobin S results from a non-synonymous mutation in the β-globin gene (HBB) on chromosome 11, which leads to valine being substituted for glutamic acid at position seven in the beta chain of adult haemoglobin. The homozygous form (HbSS) gives rise to sickle cell anaemia, and the heterozygous form (HbAS) to sickle cell trait.

The high representation of the HbS allele in some populations and correlated geographical distribution between it and malaria reflects the protection it provides against the disease (Gendrel, Kombila et al. 1991) (Carlson, Nash et al. 1994).

The malaria parasite does not survive as well in the erythrocytes of people with sickle trait as it does in the cells of normal individuals (Orjih, Chevli et al. 1985). The basis of the toxicity of sickle hemoglobin for the parasite is unknown. One possibility is that the malarial parasite produces extreme hypoxia in the red cells of people with sickle trait. These cells then sickle and are cleared (along with the parasites they harbor) by the reticuloendothelial system

(Roth, Friedman et al. 1978). Another possible mechanism is that low levels of hemichromes are formed in sickle trait erythrocytes. Hemichromes are complexes that contain heme moieties that have dissociated from the hemoglobin. Hemichromes catalyze the formation of reactive oxygen species, such as the hydroxyl radical, which can cause injury or even kill the malarial parasites (Anastasi 1984).

The malaria hypothesis maintains that during prehistory, on average, people without the sickle allele died of malaria at a high frequency. On the other hand, people with two alleles for sickle hemoglobin died of sickle cell disease. In contrast, the heterozygotes (sickle trait) were more resistant to malaria than normal individuals and yet suffered none of the ill-effects of sickle cell disease. This selection for heterozygotes is termed "balanced polymorphism" (Haldane 1949). Support for this concept comes from epidemiological studies in malaria-endemic regions of Africa. A recent study found that the state of having one sickle cell allele was associated with protection against mild clinical malaria (50%), hospital admission for malaria (75%) and severe malaria (90%). The parasite densities during clinical attacks in children with HbAS were also found to be lower than HbAA children (Williams, Mwangi et al. 2005).

Landscape of the β-globin genomic region

The β-globin gene (HBB) exists in a region of chromosome 11 called the "β-globin locus" (Figure 5.3). This is a 70kb region harboring the β-globin gene which codes the β-globin chain of adult haemoglobin, as well as related genes that code the equivalent chains during fetal life and the first few months after birth. Recombination is common around a hot spot located between the HBB and the δ-globin gene (Chakravarti, Buetow et al. 1984).

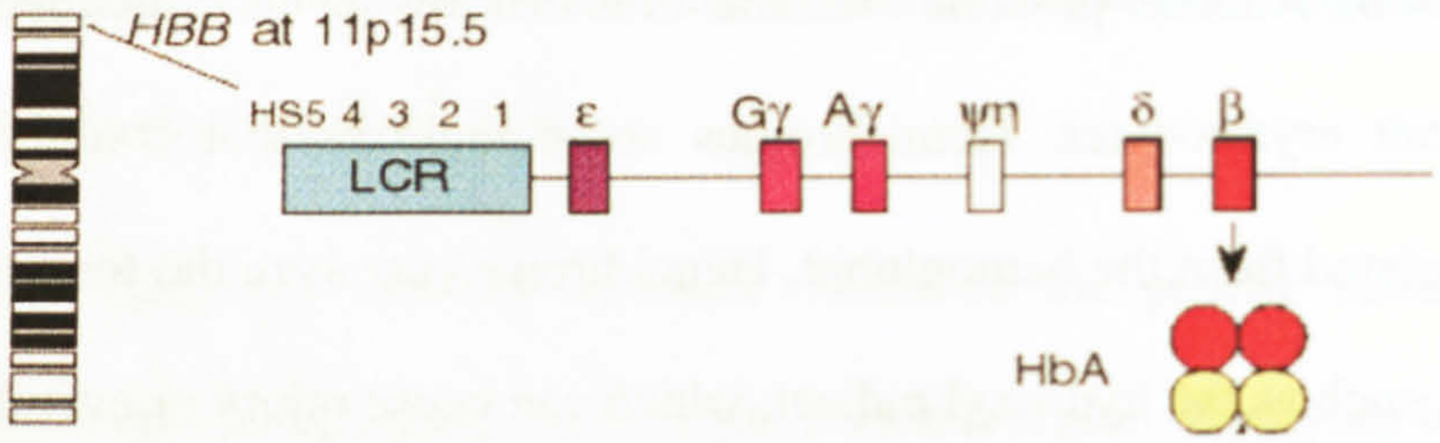


Figure 5.3a: landscape of the β-globin gene cluster on chromosome 11 p15.5 region.
The [beta]-type globin genes are clustered on chromosome 11. The figure illustrates the HBB gene which codes the β-globin chains of the adult haemoglobin. Five other *globin* genes are shown along with the locus control region (LCR), and include the embryonic [epsilon], two tandem fetal [gamma] genes, and the adult [delta] genes oriented in the 5' to 3' direction.

In the wider genomic area of 2Mb encompassing the β-globin gene cluster, there are about 23 recombination hotspots. Many of them are of a higher intensity than that near the HBB gene (Figure 5.4).

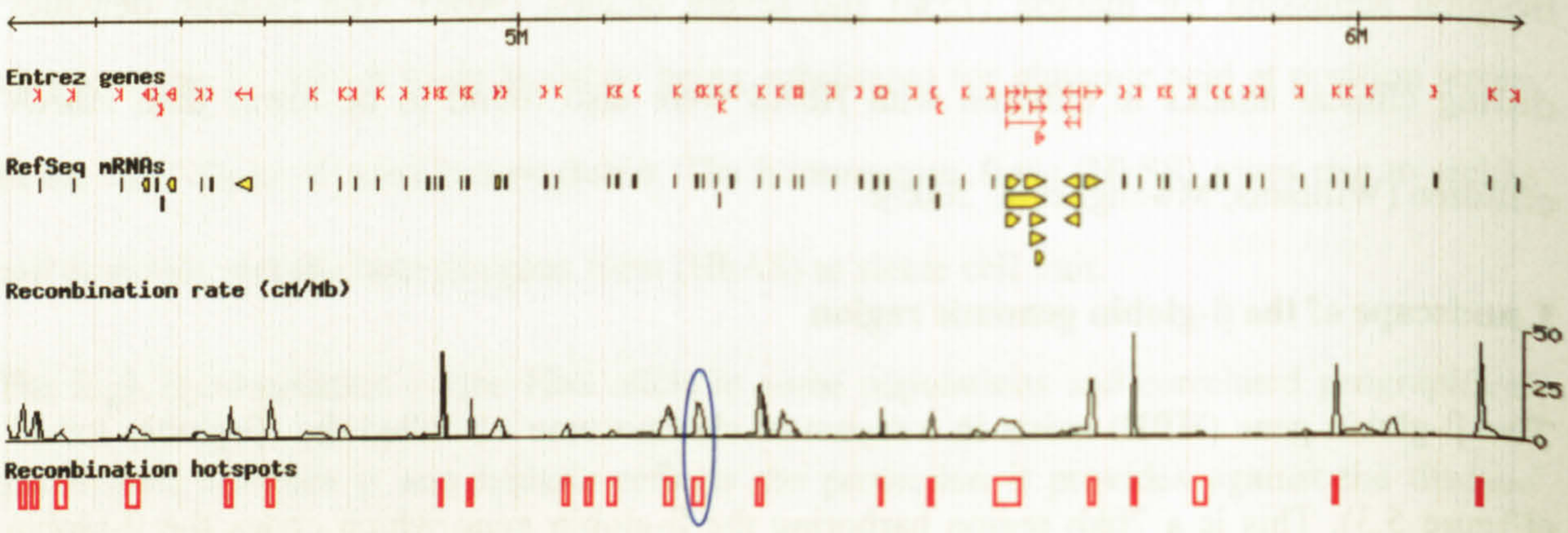


Figure 5.3b: 1.792 Mbp between position 4,397,059 and 6,189,229 of chromosome11. HapMap data release 16c.1/phase1 June05. NCBI B34 assembly. Circled in blue the hotspot near the HBB gene.

The classical HbS haplotypes

The genetic variation in the flanking regions of the HbS variant (the haplotypes) show that HbS allele arose separately at least four times in Africa, and once in Asia, possibly in India (Nagel and Fleming 1992).

The four distinct African haplotypes are all localized exclusively to a separate geographical areas, Senegal, Benin, Central African Republic (CAR), and Cameroon. The haplotypes are named for the geographic regions where they were identified. The geographic distribution of HbS haplotypes has been argued by some scientists to demonstrate the independent origin of the HbS mutation in these regions. This assumption has been rejected by others who favor a unicentric origin of the HbS mutation that would have had spread to different haplotypes by a yet to be substantiated process (Flint, Harding et al. 1993).

The four African haplotypes show broad trends in disease severity. The CAR haplotype tends to have the least favorable clinical course, followed by the Benin and Senegal haplotypes (Powars and Hiti 1993). The ranking of the more recently described fourth haplotype, Cameroon, is uncertain.

No clear explanation exists for the differences in average severity between the haplotypes. The hypothesis is that, the mutations in the flanking region could secondarily affect severity by altering fetal haemoglobin (HbF) expression in the cells. The patterns of severity apply only to populations. Broad overlap in the clinical patterns prevents the use of haplotypes to predict the clinical course in a particular person. HbF is lower in patients with Benin, CAR or Cameroon RFLP haplotypes (less than 10%) than in those with Senegal or Asian haplotypes (15-30%).

These classical globin gene cluster haplotypes are determined by DNA polymorphic sites that are identified by endonuclease enzymes. In any particular population, the majority of the

HbS chromosomes have one of the five common haplotypes, However usually 5-10% of the chromosomes of any sample have less common haplotypes, referred to as atypical haplotypes which are thought to have been generated by a variety of genetic mechanisms including (a) isolated nucleotide changes in one of the polymorphic restriction sites, (b) simple and double crossovers between two typical HbS haplotypes or much more frequently between a typical HbS haplotype and a different HbA associated haplotype that was present in the population, and (c) gene conversion (Zago, Silva et al. 2000). One study showed that 3.1% of apparently typical haplotypes involve recombinations similar to those that generate the atypical haplotypes, which reinforces the picture of the β-globin gene cluster as highly dynamic (Zago, Silva et al. 2001).

5.4. Materials and Methods

Samples

The samples for this study comprise Sudanese individuals from Hausa and Masalit tribes of Eastern Sudan. Genotyping was carried out in 48 unrelated Masalit individuals from Salala village and 47 unrelated Hausa individuals from Koka village in eastern Sudan. These samples were a subset of that genotyped previously in the 5q31 region (Chapter 3).

SEQUENOM genotyping

The ENSEMBL database (http://www.ensembl.org/) was used to identify an initial set of 28 SNPs across 414 kb of the β-globin locus on chromosome 11, spanning about 200 kb on either side of the HbS SNP. SNPs were chosen on the basis of validation (preferably in an African-related population), and available frequency data. For these markers there were already designed assays and available primers in the Kwiatkowski lab. These SNPs were selected by Neil Hanchard (D.Phil thesis 2004).

Chosen SNPs (including the HbS SNP) were genotyped using MALDI-TOF mass spectrometry (SEQUENOM) (Griffin and Smith 2000) on PEP DNA in 95 Sudanese population samples. SNPs with greater than 10% missing data, genotypes not consistent with Hardy-Weinberg equilibrium ($P < 0.01$), or minor allele frequencies < 5% were then excluded, resulting in a final set of 20 SEQUENOM-typed SNPs (SNP assay details in table 5.4.1).

SNP Assay name	rsnumber	Position	first PCR-primer sequence	second PCR-primer sequence	fragment length	UEP sequence
MMP26_ex2	rs2499953	11:4967481	ACGTTGGATGGTTTGTGTCTCCTGGGTAAG	ACGTTGGATGTTCTTGATCTGATTCAGGGC	102	CATCAATTTTTCCTGA
11-5359618	rs17497	11:5014184	ACGTTGGATGGGAACCAGTACTTTTCTGAG	ACGTTGGATGAAAAACTCTGACCCAAGCCC	105	CTCTCAGATACTGAA
OR52A1-2421	rs2472527	11:5118600	ACGTTGGATGCAACCATCCATAGTAAAAGC	ACGTTGGATGCTCCTATCTGCACGTTTTTC	118	AAAAGCAGAAACAAC
11-5469380	rs2472523	11:5123701	ACGTTGGATGGCTTGTTAATTTTCTTAGCAG	ACGTTGGATGACGACACAGGAAGCAACAAG	101	AAGATATAACCTTAT
OR51A1P_plus1941	rs4910715	11:5132596	ACGTTGGATGAACTCCTCCTACTTCTCCAC	ACGTTGGATGCACGGGATGAATTTATTGAGC	115	ATTGGAGAAAAAAGT
11-5498824	rs4910722	11:5153293	ACGTTGGATGAGACCAGTACCACTACAGAC	ACGTTGGATGTTGGTCTTTTTTCTCCCCCC	115	TGATTGATATTAACT
11-5502327	rs4910726	11:5156896	ACGTTGGATGAGGCCCTACTGGATTTAAAG	ACGTTGGATGTACCTTGATAGGCAGCATTG	102	TAGGCAGCATTGGAT
hHbS_C	rs334	11:5204808	ACGTTGGATGAAACAGACACCATGGTGCAC	ACGTTGGATGTTCATCCACGTTCACCTTGC	100	AGGGCAGTAACGGCA
HBB-703	rs11036364	11:5205580	ACGTTGGATGCTGCATTAAGAGGTCTCTAG	ACGTTGGATGCAATGTGCTCTGTGCATTAG	96	TTAGGTTTTGGGAAA
HBB-989	rs16911905	11:5205866	ACGTTGGATGAAGGAGGTTTAAACAAACAA	ACGTTGGATGAGTCCACTAAGAATACTGCG	106	GCGTTTTAAAATCAT
HBBPG1	rs2071348	11:5220722	ACGTTGGATGCTAAAATTTGGTAGAGCAAGG	ACGTTGGATGGAGCTATCAAATGGTAAGTGG	80	AAGTGGCCTTCCATT
HBG_plus6698	rs916111	11:5225919	ACGTTGGATGTTTGTGGTGTGTGCATGCTC	ACGTTGGATGAACTCCCGTGTACAAGTGTC	114	GTCTTTACTGCTTTTA
HBG-1195	rs2855122	11:5233812	ACGTTGGATGCTCTGAACTCGATCCATGAC	ACGTTGGATGAGAGCTCTCCTCCAATAAGC	97	CAGATTTCCAGAGTT
HBE-5519	rs2156918	11:5253468	ACGTTGGATGTCCTTCACAGATGCCTTAGC	ACGTTGGATGTAATTCTAGCCCCACAGGAG	102	AAAGTAAACTTCCAC
11-5620487	rs3888708	11:5274956	ACGTTGGATGCCAATGCATGCAGAAAGAAG	ACGTTGGATGGGTGACTCTAACAGAAAGGG	114	GAAAGGATGTTGGCA
11-5630468	rs3898917	11:5284937	ACGTTGGATGCTTACCTGAATATATGAGAGG	ACGTTGGATGACACTCGCAGCAAGTTACTC	107	GTTTATACTTAGGGG
OR51B2_plus6589	rs3902049	11:5294576	ACGTTGGATGATGTAGAGGTGGCATTGGAG	ACGTTGGATGTCAACCAATGTCCTCGGAAC	101	GTATTTCTCCAAGGA
11-5678383	rs2647598	11:5332852	ACGTTGGATGCAGGAAATTGGTCCAGGTAG	ACGTTGGATGAGAAAGGTGATGTGCTCTGG	110	GCATCCTTGCTAACA
11-5694576	rs728925	11:5349060	ACGTTGGATGTTTGAACTTCTGTGTTTGAG	ACGTTGGATGTAAGCCACACAGCACCTTTC	102	AGCACCTTTCCCCGG
OR51JIP	rs872165	11:5381108	ACGTTGGATGACCGTATTGGGTATTGCCAC	ACGTTGGATGTAAGCAAGCACAGCACAGAC	103	TTGAGTGCCTTCTTC

Table 5.4.1: β-globin PCR primers.
Assay details of those SNPs typed in the β-globin region using SEQUENOM. 1st and 2nd round PCR Mass Spectrometry primers used for HBB genotyping are shown.
Also included are the rs reference numbers as well as their location on chromosome 11 (Ensembl release 50).

A subset of the samples, namely those found to be carrying the sickle cell allele (12 Masalit and 9 Hausa), were further genotyped for another 37 markers spaced across a genomic area measuring about 2 Mb around the HbS polymorphism (see table 5.4.2. for details).

MARKER ID	SNP ASSAY NAME	RS NMBER	POSITION
1	11-4397059	rs10500600	11:4397059
2	11-4440254	rs11032345	11:4440254
3	11-4453331	rs2278170	11:4453422
4	11-4507222	rs11032629	11:4507222
5	11-4521274	rs1505212	11:4521274
6	11-4528237	rs12281831	11:4528237
7	11-4682971	rs16933304	11:4682971
8	11-4755172	rs2898966	11:4755087
9	11-4755524	rs1594814	11:4755524
10	11-4800616	rs17328191	11:4800616
11	11-4842124	rs2196122	11:4842124
12	11-4857919	rs16906893	11:4857919
13	11-4877465	rs16907096	11:4877465
14	11-4911118	rs10500623	11:4911118
15	11-5070787	rs1551489	11:5070787
16	rs7114854	rs7114854	11:5100498
17	11-5519408	rs4910732	11:5173877
18	hHbC_B	rs33930165	11:5204809
19	11-5206744	rs7936823	11:5206744
20	11-5207406	;	11:5207406
21	11-5207723	;	11:5207723
22	11-5207734	;	11:5207734
23	11-5254606	rs10488675	11:5254606
24	11-5313624	rs11036885	11:5313624
25	rs7938837	rs7938837	11:5319683
26	rs7929631	rs7929631	11:5324552
27	11-5343364	rs10500637	11:5343364
28	rs1498468	rs1498468	11:5367607
29	11-5502292	rs317776	11:5502292
30	11-5632519	rs16933926	11:5632519
31	11-5667472	rs2291841	11:5667472
32	11-5676411	rs1063303	11:5676326
33	11-5929083	rs4453215	11:5929083
34	11-6111354	rs325632	11:6111354
35	11-6180413	rs188980	11:6180413
36	11-6181392	rs4758398	11:6181392
37	11-6189229	rs9659	11:6189229

Table 5.4.2: Extra SNPs genotyped in subset of Sudanese samples found to be carrying HbS allele.
Listed are the rs reference numbers as well as the location of SNPs on chromosome 11 (ENSEMBL release 39). SNPs number 20, 21 and 22 were chosen from literature and did not have an rs entry in the public database (Wood, Stover et al. 2005).

RFLP genotyping

Six RFLP sites were additionally genotyped in the same 95 Sudanese population samples. The RFLP markers were chosen so as to characterize the classical HbS haplotypes in the 70 kb β-globin-like cluster region on chromosome11. Using PCR amplification and subsequently digestion with restriction endonuclease enzymes, the restriction enzymes sites Hinf I, Hinc II, Hind III in HBG1, Hind III in HBG2, and Xmn I were typed.

Hinf I digests were uninformative as there were multiple Hinf I sites in the amplified PCR fragment; the remaining restriction digests were therefore used to define classically-described β^s haplotypes.

HBB PCR Primers (see table 5.4.3) used to amplify products for RFLP genotyping were designed with careful consideration of the high degree of homology in the region due to gene duplication. Primers were designed to make sure they amplify a unique segment containing the targeted markers. This resulted in relatively large amplification fragments. The HBG2 fragment (2734 bp in length) amplified the HBG2 gene and contained restriction sites for both Hind III and Xmn1. The HBG1 fragment (2909 bp in length) amplified the HBG1 gene and contained the restriction site Hind III. The HBB fragment (1200 bp in length) amplified the HBB gene and contained the restriction site Ava II. The recognition site for Hinc II was in an intergenic region with unique flanking sequence, so a small fragment of 118 bp containing it was amplified.

Restriction enzyme	rs number	Position	first PCR-primer sequence	second PCR-primer sequence	fragment le
Ava II	rs10768683	5204367	AAATTAAGAAAAACAACAACAAATGAATG	CATTCTAAACTGTACCCTGTTACTTATCC	1200
Hinf 1	rs10742584	5205346	AAATTAAGAAAAACAACAACAAATGAATG	CATTCTAAACTGTACCCTGTTACTTATCC	1200
Hinc II	rs968857	5217034	ACGTTGGATGTCTGCCTCTGCTATAGTCTG	ACGTTGGATGCTGACTTCTGATACTATGTCT	118
Hind III	rs6578593	5226375	ACGTTGGATGCATGTACACGCACATCTTATGTC	ACGTTGGATGCTTAAGAACCATCCTTGCTACTCAG	2909
Hind III	rs2070972	5231293	GACAGCATGAATACTTCCTGCC	ACGTTGGATGGAACTGAAGACAACCATGTGTG	2734
Xmn 1	rs7482144	5232745	ACGTTGGATGACAGCATGAATACTTCCTGCC	ACGTTGGATGGAACTGAAGACAACCATGTGTG	2734

Table 5.4.3: Positions of RFLP markers and primer sequences used to amplify PCR products in the HBB region.
Lengths f the amplified fragments are shown.

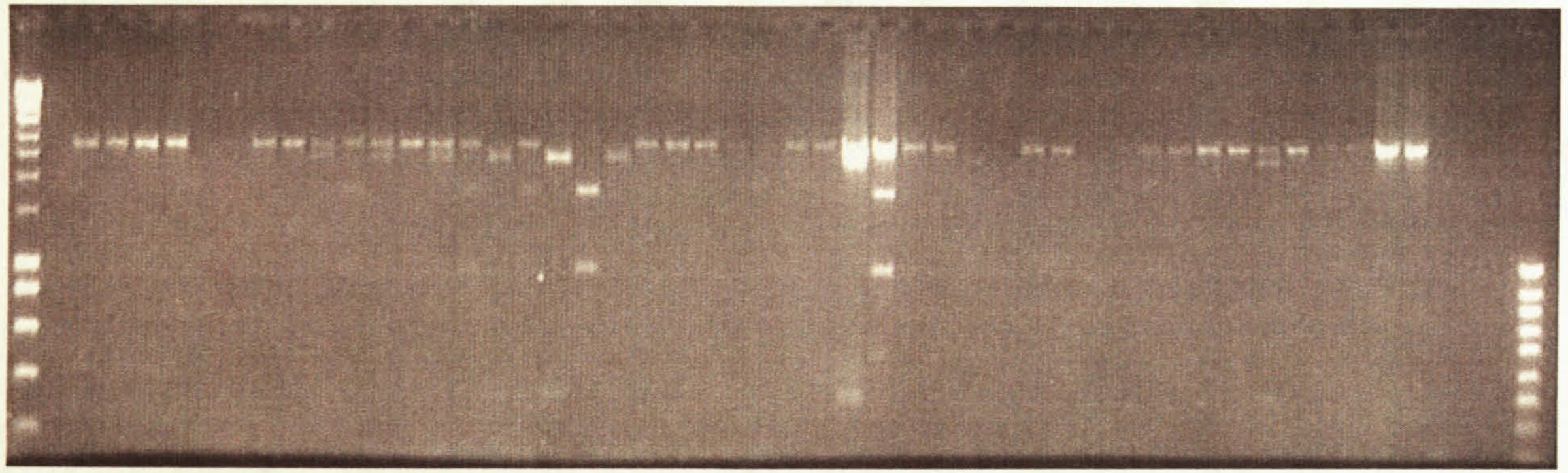


Figure 5.4: An agarose gel electrophoresis image of digestion products of The HBG2 fragment.
Loaded in alternating wells on the gel are digestion products of HindIII and XmnI. The sizes of the digested fragments indicate the genotype at the enzyme recognition site.

For each PCR reaction 2 μL of genomic DNA at a concentration of 5 ng/μL was added to 6 μL of PCR mix. PCR mix for 192 reactions was prepared by adding the following: $MgCl_2$ (50 mM) – 44 μL; dNTPs (8 mM pool) – 110 μL; ×10 buffer – 110 μL; Biotag 5 U/μL – 5.5 μL; H_2O – 386.1 μL; 1st PCR primer – 2.2 μL; 2nd PCR primer – 2.2 μL. The PCR mix was the same for all fragments except HBG1 (2909 bp) for which 3.3 μL of each of the forward and reverse primers was used.

PCR protocols for the HBG1 and HBG2 RFLP fragments consisted of an initial five cycle denaturation of 96°C for 1 minute, 94°C for 45 seconds, 62°C for 2.5 minutes, and 72°C for 1 minute; followed by a 29 cycles of 94°C for 45 seconds, 65°C for 2.5 minutes, and 72°C for 1 minute, and a final extension of 72°C for 10 minutes and 15°C for 15 minutes. The PCR protocol for the Hinc II fragment differed only with regard to the main cycling conditions which required an annealing temperature of 65°C for 45 seconds and an extension temperature of 72°C for 30 seconds. The HBB fragment did not require the initial 5 cycle denaturation; instead 35 cycles consisting of 96°C for 1 minute, 94°C for 45 seconds followed by an annealing temperature of 56°C for 45 seconds, and a 72°C extension for 1 minute was used (For full PCR amplification and digestion protocols see Materials and Methods, chapter2).

Restriction enzymes and their buffers were ordered from New England BioLabs (Ipswich, MA, USA); digests were carried out according to the manufacturer's recommendations. Digestion products were loaded onto an agarose gel and scored as +/+ if the two alleles were digested, as +/- if one but not the other allele was digested (heterozygote), and as -/- if no digestion occurred in the sample (Figure 5.4).

Analytical and statistical methods:

Fst

Wrights Fixation Index statistic (Fst), which is a measure of inter-population diversity that uses the difference between the average observed and the total expected heterzygosity; was calculated for each of the SNPs typed.

STRUCTURE

Both the basic model and the linkage model were used in the software ***STRUCTURE*** on the genotypic data and haplotypic data respectively from the two study populations.

The program ***STRUCTURE*** implements a model-based clustering method for inferring population structure using genotype data. A model in which there are K populations (where K may be unknown) is assumed. Each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations. It is assumed that within populations, the loci are in Hardy-Weinberg equilibrium, and linkage equilibrium. Individuals are assigned to populations in such a way as to achieve this.

Long-range haplotype similarity

The Sudanese haplotypes (each population separately) in the HBB genomic region, were uploaded into **MARKER** (www.gmap.net/marker), and the HAPLOSIMILARITY algorithm was used to calculate the haplotype similarity (HS) scores. This algorithm uses sliding windows to assess the mean similarity of haplotypes (given as the mean of the sum of the squares of the frequencies of distinct haplotypes within a given window) associated with the minor allele of a given SNP. The value of haplosimilarity ranges from one (all haplotypes associated with the allele are exactly the same) to a minimum given by $1/k_{max}$, where k_{max} is the maximum possible number of distinct haplotypes for a given sliding window size (haplotypes associated with the allele are extremely diverse). I used the default option for a sliding window size, which is ten SNPs, in my evaluation.

HAPLOSIMILARITY (including details on operating characteristics and implementation) is available for public use at the GMAP website (http://www.gmap.net/pub/003).

Extended Haplotype Homozygosity (EHH)

The EHH statistic of the LRH, implemented in the software *Sweep*, is broadly similar to the haplosimilarity statistic and is defined as the probability that at a given distance away from a core haplotype or SNP, any two haplotypes extending outward from the core haplotype/SNP will be homozygous at all SNPs. EHH scores range from a minimum of zero to a maximum of one.

The EHH reflects the rate of decay in LD at increasing distance from a locus and uses this to determine the age of the relevant allele. Core haplotypes are defined to mark subsets of similar haplotypes in the region. The EHH of a given haplotype subset, marked by its core haplotype, is compared to other core haplotypes in the population. The relative EHH (rEHH) of a haplotype refers to its EHH relative to the EHH of other core haplotypes. The inclusion of alternative haplotypes as controls for one another is advantageous because it intrinsically controls for potential confounders due to variation in the local recombination rate (Sabeti, Reich et al. 2002).

5.5. Results

Genotyping success rate was above 92% for all of the typed markers in the HBB region. All genotyped markers were found to be in Hardy-Weinberg equilibrium.

5.5.1. Allele Frequencies in the two populations

The allele frequencies of genotyped markers were found to be correlated between the two populations (figure 5.5.1.1 and figure 5.5.1.2). This result is similar to that previously seen in the 5q31 region.

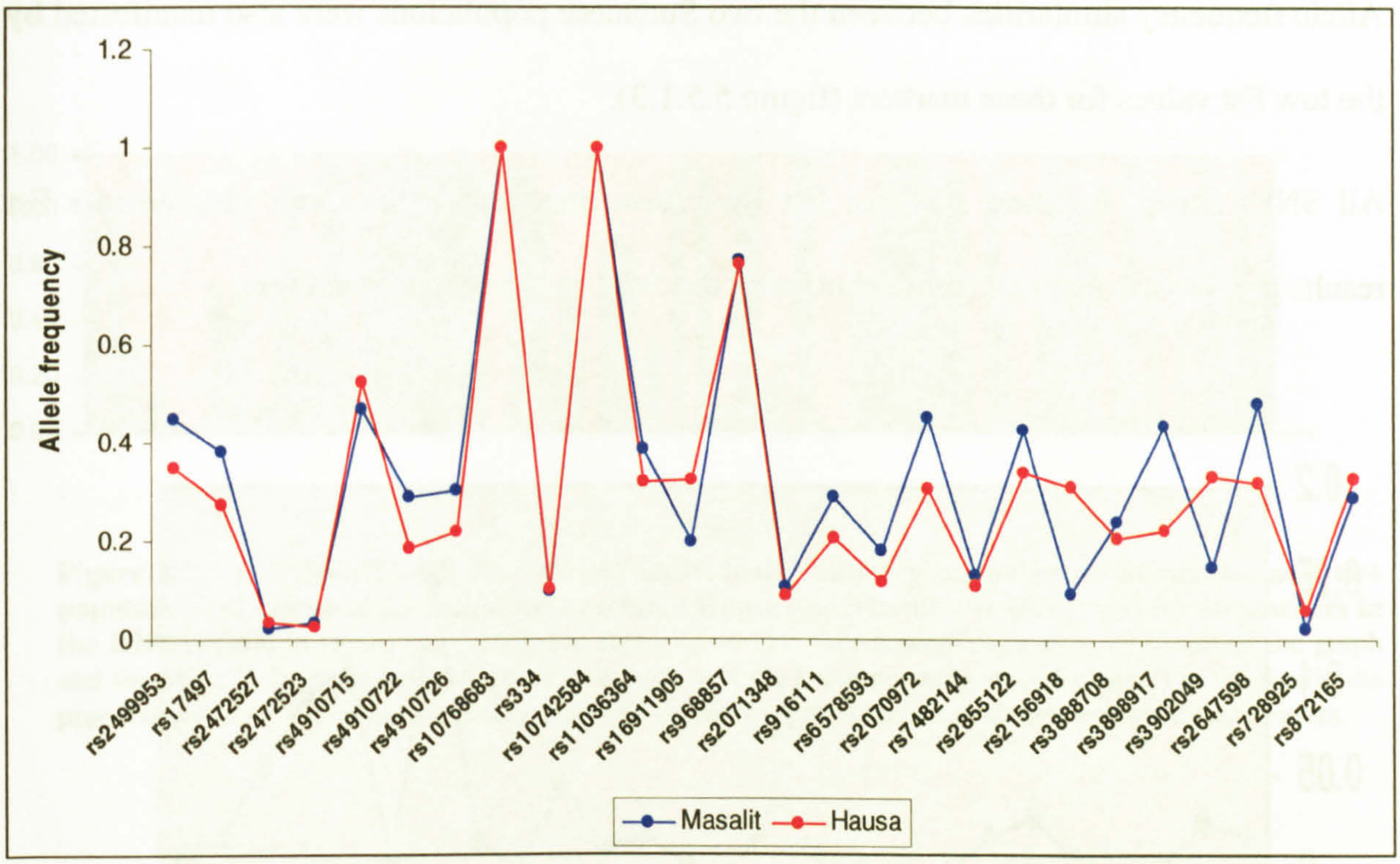


Figure 5.5.1.1: Allele frequencies of markers genotyped in the HBB in the Sudanese Hausa and masalit samples. Markers are shown on the x axis by their rs numbers and allele frequencies on the y axis. Red points represent allele frequency values in Hausa. Blue points represent those in the Masalit.

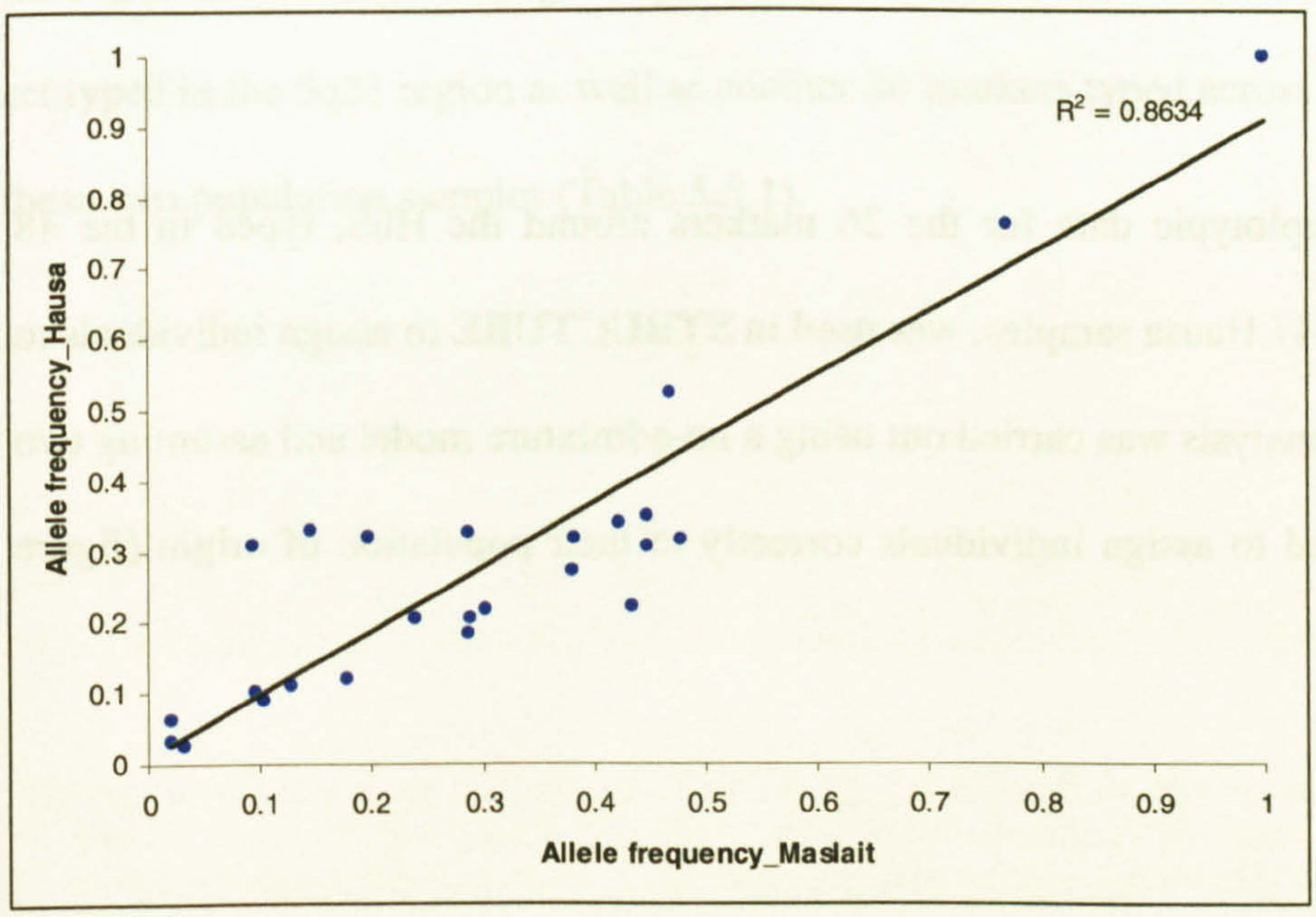


Figure5.5.1.2: The correlation of allele frequencies between the Hausa and Masalit samples. Shown is data for 26 markers in a 400kb region around the HbS variant. The correlation coefficient R^2 is shown on the top right corner of figure.

Allele frequency similarities between the two Sudanese populations were also manifested by the low Fst values for these markers (figure 5.5.1.3).

All SNPs except for three have an Fst value less than 0.06. This result also mirrors Fst results previously obtained from genotyping the 5q31 in the two populations.

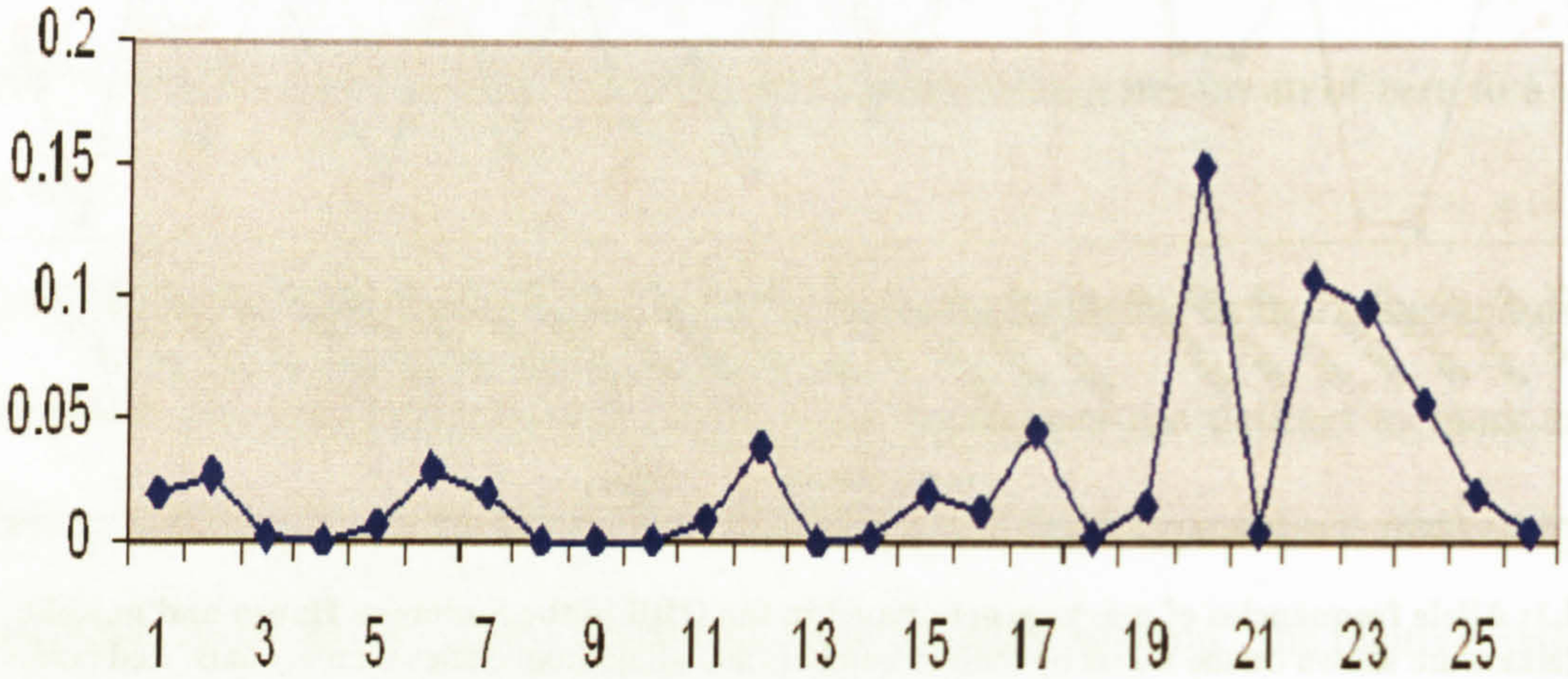


Figure 5.5.1.3: Single-SNP Fst values for markers typed in the HBB region in the Hausa and Masalit samples. Fst values are shown in the y axis and markers ordered on the x axis by position as shown in figure 5.5.1.1.

The genotypic and haplotypic data for the 26 markers around the HbS, typed in the 48 unrelated Masalit and 47 Hausa samples, was used in **STRUCTURE** to assign individuals to their population. The analysis was carried out using a no-admixture model and assuming two populations. This failed to assign individuals correctly to their population of origin (figure 5.5.1.4).

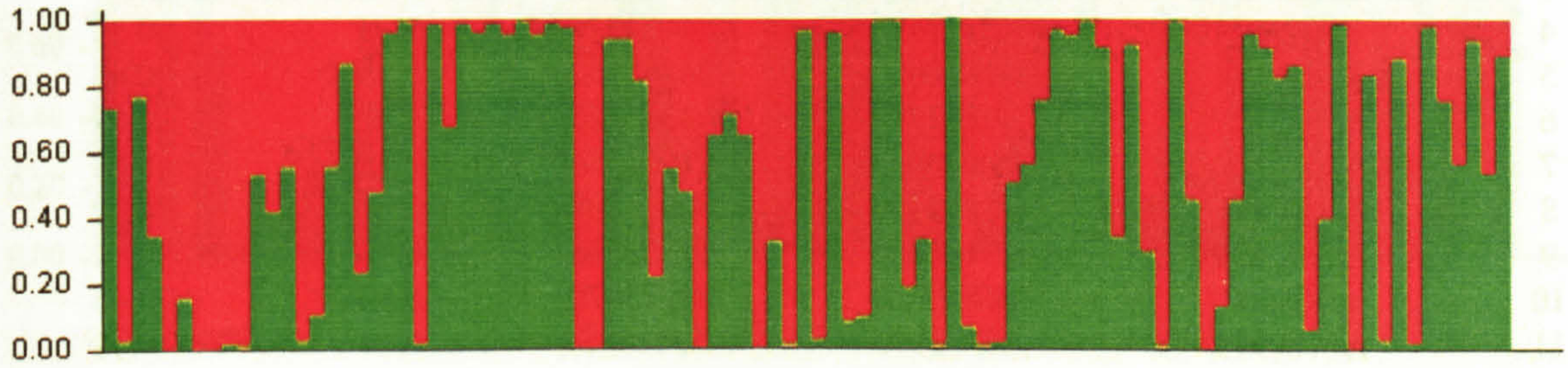


Figure 5.5.1.4: STRUCTURE Bar plot of individuals' ancestry assuming no admixture and two populations of origin of the combined unrelated Hausa and Masalit samples typed for 26 markers in the HBB region. In figure individuals are arranged so that Hausa sample constitute left half of the graph and the Masalit the right half. On the x axis each vertical bar represents an individual. On the y axis the proportions of the individuals ancestry assigned to the two populations are shown with different colours.

Individuals were not identifiable by their genetic variation data at the typed loci as belonging to their sampled geographic/population assignment.

This does not seem to be a region-specific effect. This result was mirrored in the 29-marker set typed in the 5q31 region as well as another 36 markers typed across the whole genome of these two population samples (Table 5.5.1).

id	Assay name	Assay (rs number)	coordinate	Hausa MAF	Masalit MAF
1	rs7537937	rs7537937	1:89355278	0.45	0.39
2	rs1801274	rs1801274	1:159746369	0.50	0.42
3	IL10_232424450	rs3024500	1:205007454	0.46	0.49
4	12-72249510	rs2227491	12:66932788	0.33	0.34
5	hIL-10-1082	rs1800896	1:205013520	0.38	0.43
6	rs9282799	rs9282799	17:23152855	0.07	0.07
7	hICAM-1codon29	rs5491	19:10246540	0.15	0.08
8	hIL-10-3533	rs1800890	1:205015988	0.24	0.22
9	rs1143634	rs1143634	2:113306861	0.13	0.12
10	rs708567	rs708567	3:9935070	0.50	0.45
11	rs6780995	rs6780995	3:57113459	0.44	0.49
12	hTNF-308	rs1800629	6:31651010	0.05	0.04
13	hTNF-238	rs361525	6:31651080	0.06	0.04
14	hTNF_plus851	rs3093662	6:31652168	0.08	0.07
15	rs1555498	rs1555498	6:137367540	0.46	0.48
16	hCD36_T1264G	rs3211938	7:80138385	0.17	0.00
17	rs8176747	rs8176747	9:135121136	0.22	0.16
18	rs229587	rs229587	14:64333053	0.40	0.49
19	rs8176746	rs8176746	9:135121143	0.23	0.16
20	12-72251611	rs2227478	12:66934889	0.36	0.43
21	IL4R-63011	rs1805015	16:27281681	0.48	0.33
22	hNOS2-1659	rs8078340	17:23153339	0.33	0.23
23	rs17047661	rs17047661	1:205849512	0.28	0.26
24	rs17411697	rs17411697	2:113259694	0.14	0.13
25	hICAM-1codon469	rs5498	19:10256683	0.08	0.02
26	rs11096957	rs11096957	4:38452886	0.44	0.39
27	hIL-4-589	rs2243250	5:132037053	0.24	0.15
28	hLT-alpha_NcoI	rs909253	6:31648292	0.52	0.43
29	hTNF-376	rs1800750	6:31650943	0.04	0.03
30	12-72245636	rs2227507	12:66928914	0.00	0.04
31	12-72247607	rs1012356	12:66930885	0.49	0.35
32	hNOS2-954	rs12720463	17:23152636	0.04	0.05
33	12-72250702	rs2227485	12:66933980	0.45	0.36
34	rs77806	rs77806	14:64322985	0.45	0.48
35	rs3177244	rs3177244	22:22509132	0.49	0.36
36	rs8176743	rs8176743	9:135121236	0.21	0.16

Table 5.5.1: Extra markers typed across the genomes of Hausa and Masalit. Shown are rs numbers of the typed assays, their chromosomal locations and their Minor Allele frequencies in the Hausa and Masalit samples.

Using a 91 Marker data set (26 in the HBB region, 29 in the 5q31 region and 36 across the genome of 48 Masalit and 47 Hausa), ***STRUCTURE*** failed to assign individuals to their self-specified populations of origin (figure 5.5.1.5). This observation suggests that rather than being a mere marker-resolution issue of the data set, this result might reflect real genetic similarities between the two populations.

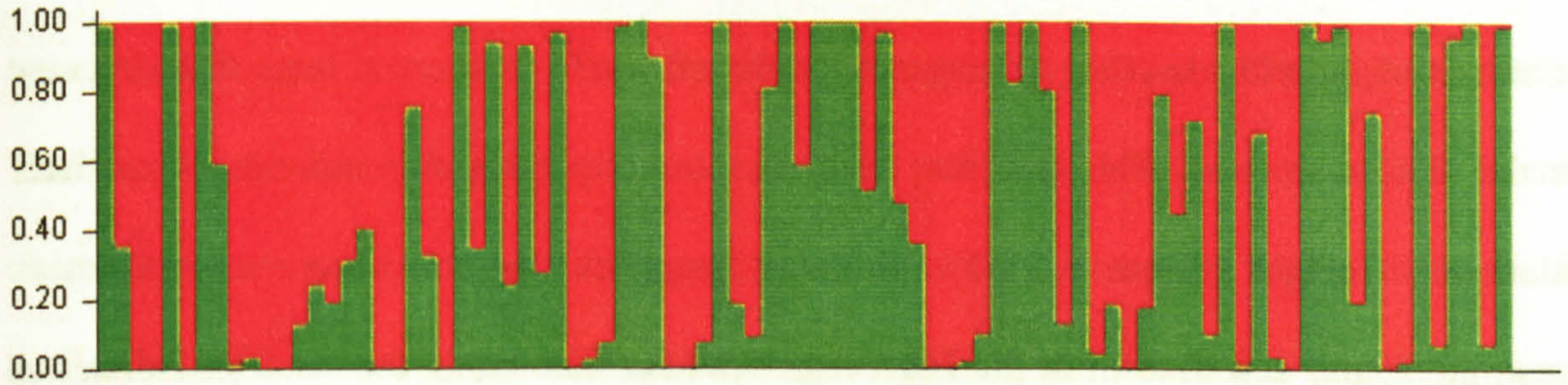


Figure 5.5.1.5: STRUCTURE Bar plot of individuals' ancestry assuming no admixture and two populations of origin of the combined unrelated Hausa and Masalit samples. Analysis was carried out using all available genotype data from 92 markers. In figure individuals are arranged so that Hausa sample constitute left half of the graph and the Masalit the right half. On the x axis each vertical bar represents an individual. On the y axis the proportions of the individuals ancestry assigned to the two populations are shown with different colours.

Interestingly, when only the subset of Hausa and Masalit individuals carrying the HbS variant were analysed by the program ***STRUCTURE***, individuals (except in a few cases) were distinctly clustered to their known population of origin (figure 5.5.1.6). This analysis was carried out with the 26-marker dataset in the 400 kb region around the HbS marker.

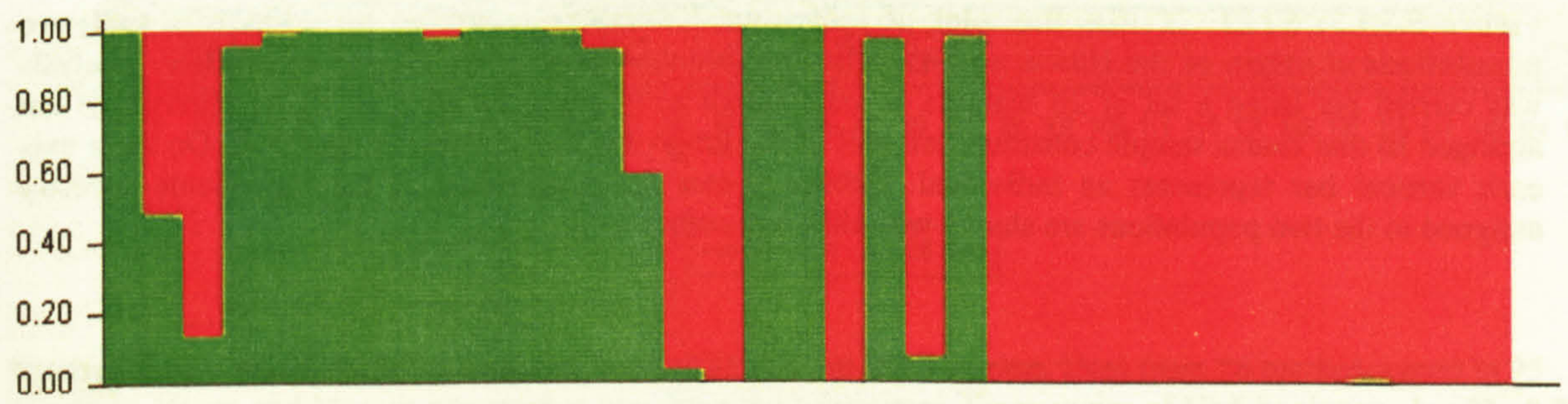


Figure 5.5.1.6: STRUCTURE Bar plot of individuals' ancestry of the Hausa and Masalit individuals heterozygote for the HbS allele. Analysis was carried out using the data for the 26 markers typed in the HBB region, and assuming no admixture and two populations of origin. In figure individuals are arranged so that Hausa sample constitute left half of the graph and the Masalit the right half. On the x axis each vertical bar represents an individual. On the y axis the proportions of the individuals ancestry assigned to the two populations are shown with different colours.

In order to exclude that the above observation is not the result of sample choice, I used the same subset of individuals. This time, genotype data from 26 markers at least 100 kb away from the HbS was used. This did not result in correctly assigning individuals to their populations of origin (figure 5.5.1.7). This fact suggests that the observed similarities between the Hausa and Masalit in the HBB region, rather than being the result of balancing selection acting on the HbS simultaneously in the two populations, is a reflection of their genome wide similarity. Furthermore, this similarity is observed in the HBB region in spite of selection sweeping the different HbS haplotypes up in the two populations, with whatever alleles happened to be on those haplotypes, leading to a greater divergence among the HbS haplotypes under positive selective pressure.

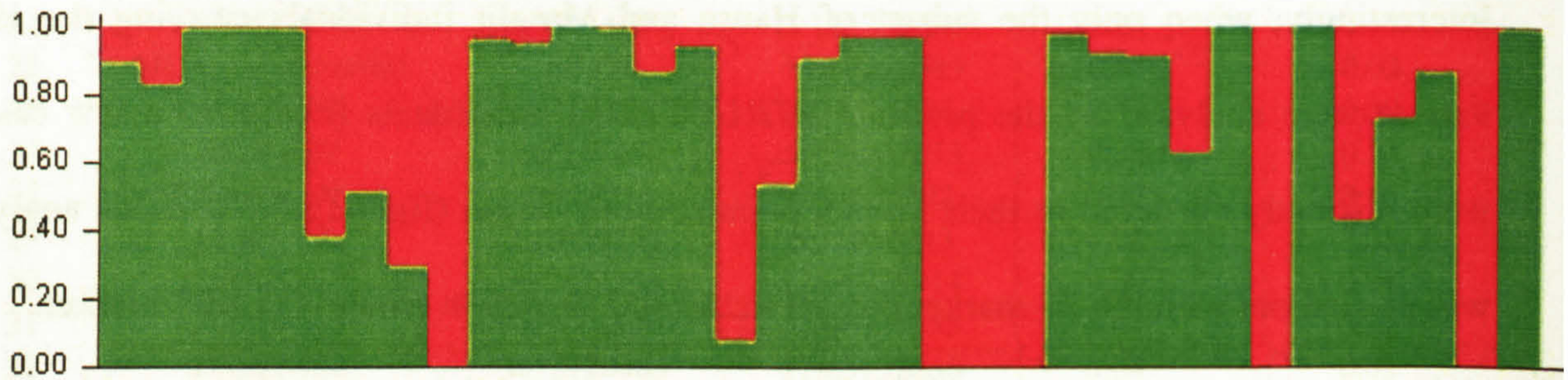


Figure 5.5.1.7: STRUCTURE Bar plot of individuals' ancestry assuming no admixture and two populations of origin of the Hausa and Masalit individuals heterozygote for the HbS allele. Analysis was carried out using a set of 26 markers at least 100kb away from the HbS. In figure individuals are arranged so that Hausa sample constitute left half of the graph and the Masalit the right half. On the x axis each vertical bar represents an individual. On the y axis the proportions of the individuals ancestry assigned to the two populations are shown with different colours.

5.5.2. Haplotype analysis

Phasing the genotypic data from the combined Hausa and Masalit sample (95 individuals) was carried out using the program **PHASE** *2.1* (see appendix 2 for full list of haplotypes, their sequences and **PHASE** probabilities).

Studying the frequency of haplotypes spanning the 400 kb region surrounding the HbS variant, haplotype frequencies were found to be generally low over this physical distance. The vast majority of haplotypes in the region had only one or two copies (identical chromosomes with the same allele sequence). The only exception to that was the HbS haplotypes. There were six identical HbS haplotypes in the combined population sample which were the most frequent across the region (Figure 5.5.2.1).

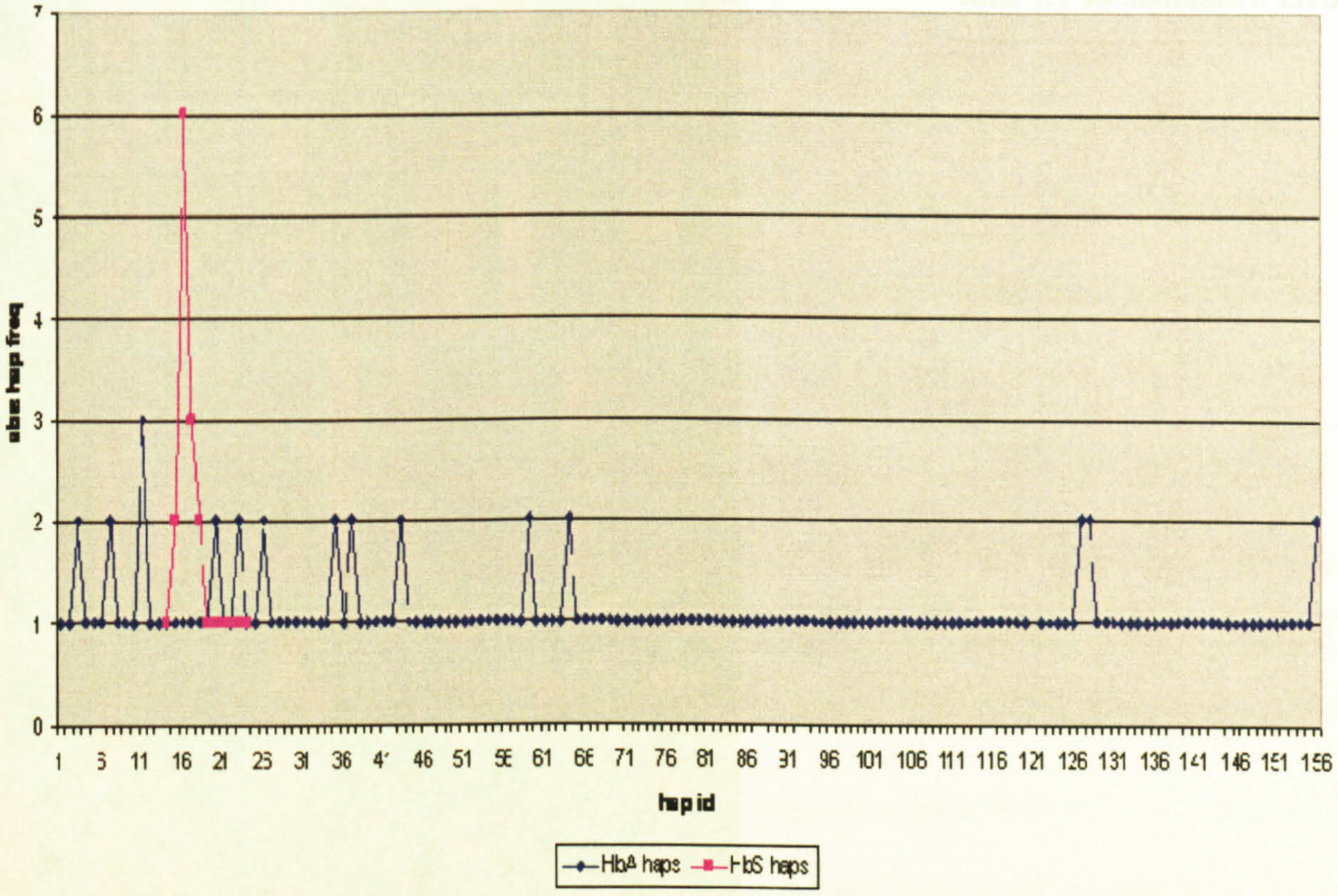


Figure5.5.2.1: Absolute frequencies of haplotypes in the HBB region. Data from 26 markers typed in 95 unrelated Hausa and Masalit was used to construct the haplotypes. Frequencies of HbS haplotypes are shown in purple. Shown on the x axis are the ids of the distinct haplotypes in the sample, and on the y axis the number of copied of each of these distinct haplotypes (see appendix 2 for full list of haplotypes, their sequences and phase probabilities).

Haplotype sequences in the HBB region are shown in Figure 5.5.2.2 and Appendix 2. Another aspect of the data that seems to differentiate the HbS haplotypes from the rest is that the haplotypic background of the HbS allele appears to be generally more homogenous than that of the HbA haplotypes (Figure 5.5.2.2).

With further analysis of the haplotypes carrying the HbS allele, and using data from the extra 37 markers characterizing the 2 Mb region anchored on the HbS variant (table 5.4.2); it was evident that the six high frequency HbS-bearing haplotypes maintained the same frequency over a distance of 1.7 Mb.

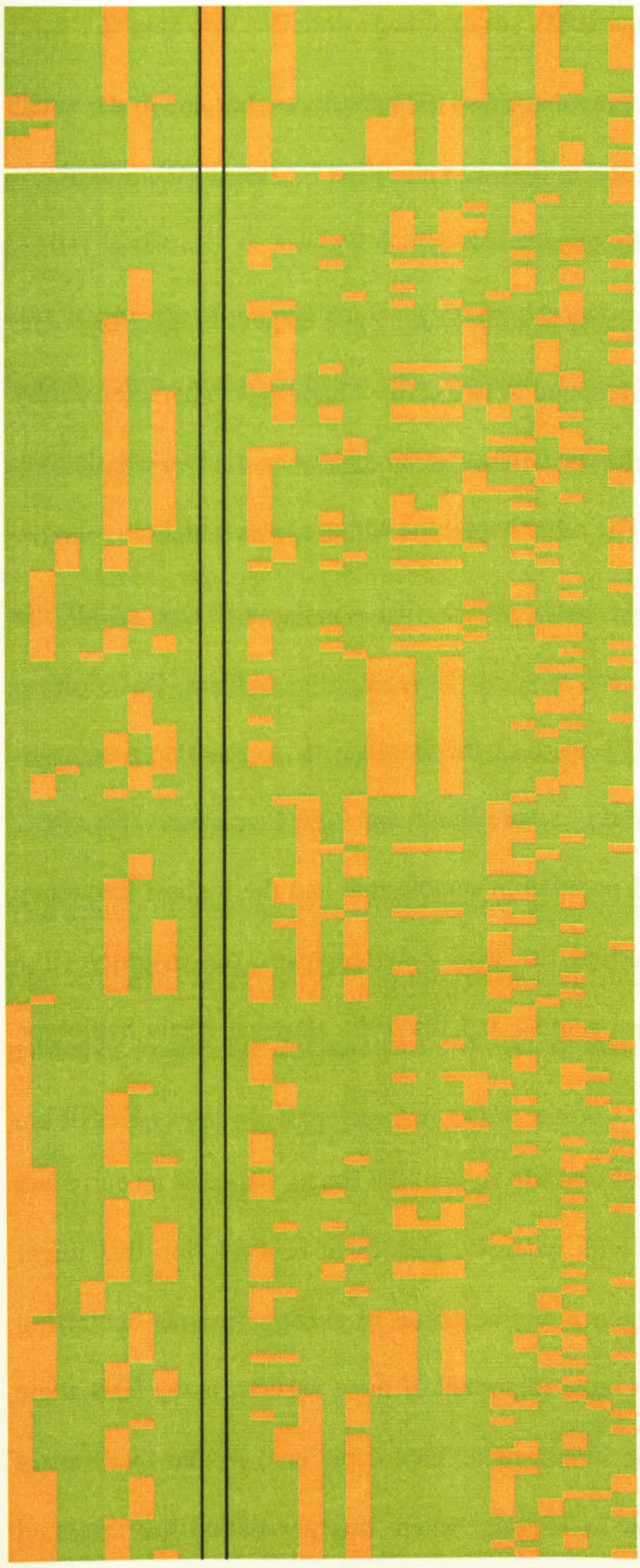


Figure5.5.2.2: Haplotypes in the HBB region in the combined Sudanese sample. Haplotypes carrying the HbS are shown in part A; HbA haplotypes are shown in part B. HbS marker position is outlined by the black border. Haplotypes are arrayed along the Y-axis (see appendix 2 for full list of haplotypes, their sequences and phase probabilities). 26 SNPs are displayed on the X-axis oriented in the 3' to 5' direction from left to right (For names and order of markers see figure 5.5.1.1). At each SNP position, the major allele of each SNP is represented in Green and the minor allele in orange.

The HbS haplotypes in Hausa and Masalit

The program **PHASE** *2.1* was initially used to partition the haplotypes, but due to the small sample size and low phase probabilities, the output was modified for the HbAS individuals so as to favor the construction of one of the classical HbS haplotypes.

There were 17 individuals from the Hausa and Masalit sample that were HbS heterozygotes. The 17 HbS haplotypes were found to correspond to two distinct haplotypes characterized by the RFLP classical markers. The first haplotype which is shared between the two populations fits the Benin classical sickle haplotype. The other haplotype which agrees with the Cameroonian haplotype was also shared between the two populations. Out of all the haplotypes of the Masalit sample 9.4% were carrying the HbS polymorphism. Out of those 78% were of the Cameroon type and 22% were of the Benin type. In the Hausa samples 8.5% of all haplotypes were HbS. 75% of those were Benin and 22% Cameroon. Out of the six identical haplotypes in the combined population sample that had the highest frequency among all the other haplotypes and carried the HbS polymorphism, the majority (five haplotypes) was contributed by the Hausa sample and fitted the classical Benin haplotype when interrogating its RFLP markers.

The recent West African origin and agricultural life style of the Hausa, suggests an early and stable exposure to the malaria parasite, with extended periods of co-evolution that might have allowed enough time for the emergence and selection of genetic variants conferring resistance to malarial disease. One of the more important of these variants is the HbS allele which is present in the Hausa on a more homogenous haplotypic background that mostly resembles the Benin haplotype dominant in Nigeria, where this population has migrated from.

5.5.3. LD map and selection signals in the HBB region

Using the programme *Sweep*, EHH was calculated for the 26 SNPs genotyped in the HBB region in the 95 unrelated Hausa and Masalit sample. The extended haplotype homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent (as assayed by homozygosity at all SNPs) for the entire interval from the core region to a distance *x* (Sabeti, Reich et al. 2002).

Core haplotypes were defined as single SNPs by setting the core selection function in the program to look at single SNPs.

To compare the EHH value between SNPs across the region, a distance measure was chosen to match those values at 0.4 cM, because it is more relevant to compare across genetic distance than physical distance. The program employs the fine-scale recombination map based on the program LDHat for the HapMap (McVean, Myers et al. 2004).

The scatter plot below gives EHH plotted against frequency for every core SNP in the data files of the 26 Markers typed in the Hausa and Masalit samples. EHH values are given at a particular long-range distance (0.4 cM). As shown in the figure, HbS had the highest EHH and REHH values of all the markers in the region.

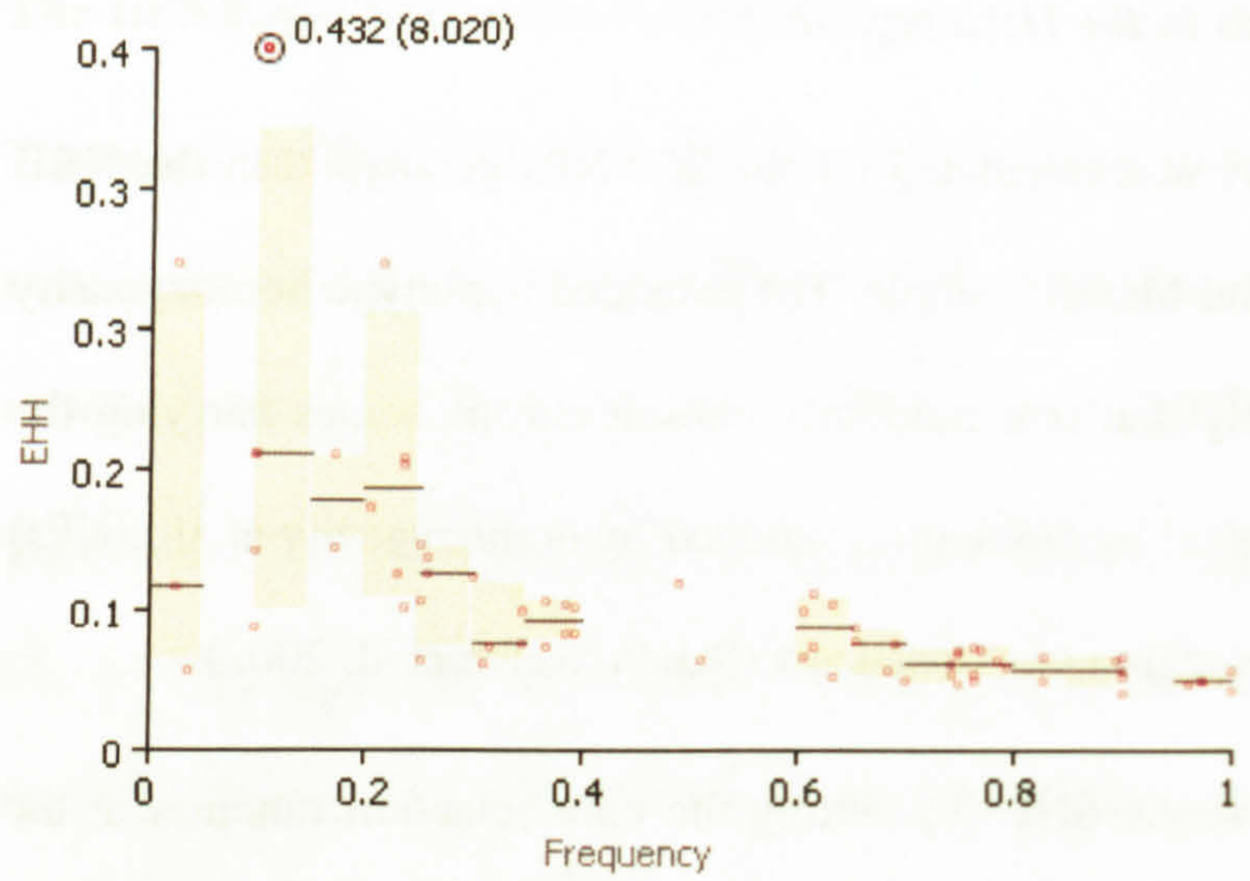


Figure 5.5.3.1: *Sweep* EHH vs. Frequency Scatter plot of the HBB region.
EHH values for every SNP (*y* axis) is shown against its allele frequency (*x* axis). The HbS marker is circled and its EHH(REHH) values are indicated in the figure.

The chart below gives EHH plotted for the HbS core SNP at every long-range distance in both directions. The different haplotypes associated with each of the SNP alleles are shown with different color in the plot.

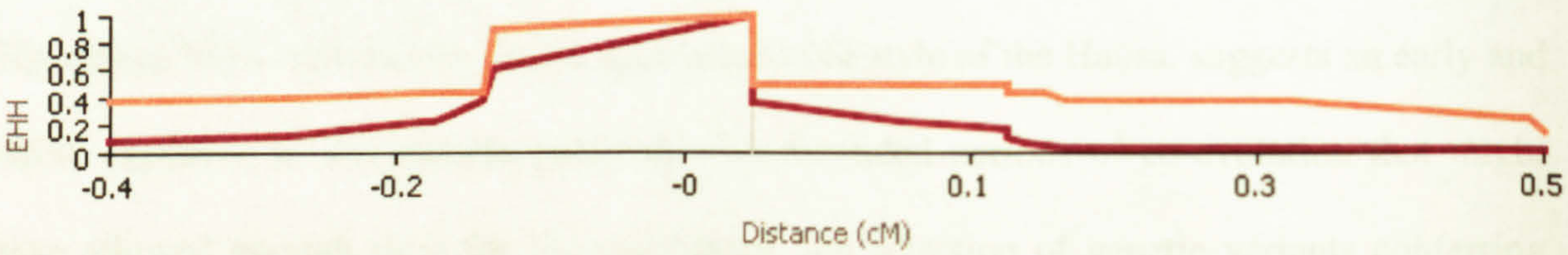


Figure 5.5.3.2: *Sweep* EHH vs. Distance Chart. EHH values are displayed on the y axis at consecutive genetic distances away from the HbS core in both directions (x axis). The haplotype carrying the HbS allele is shown in orange and the haplotype with the HbA allele is shown in purple.

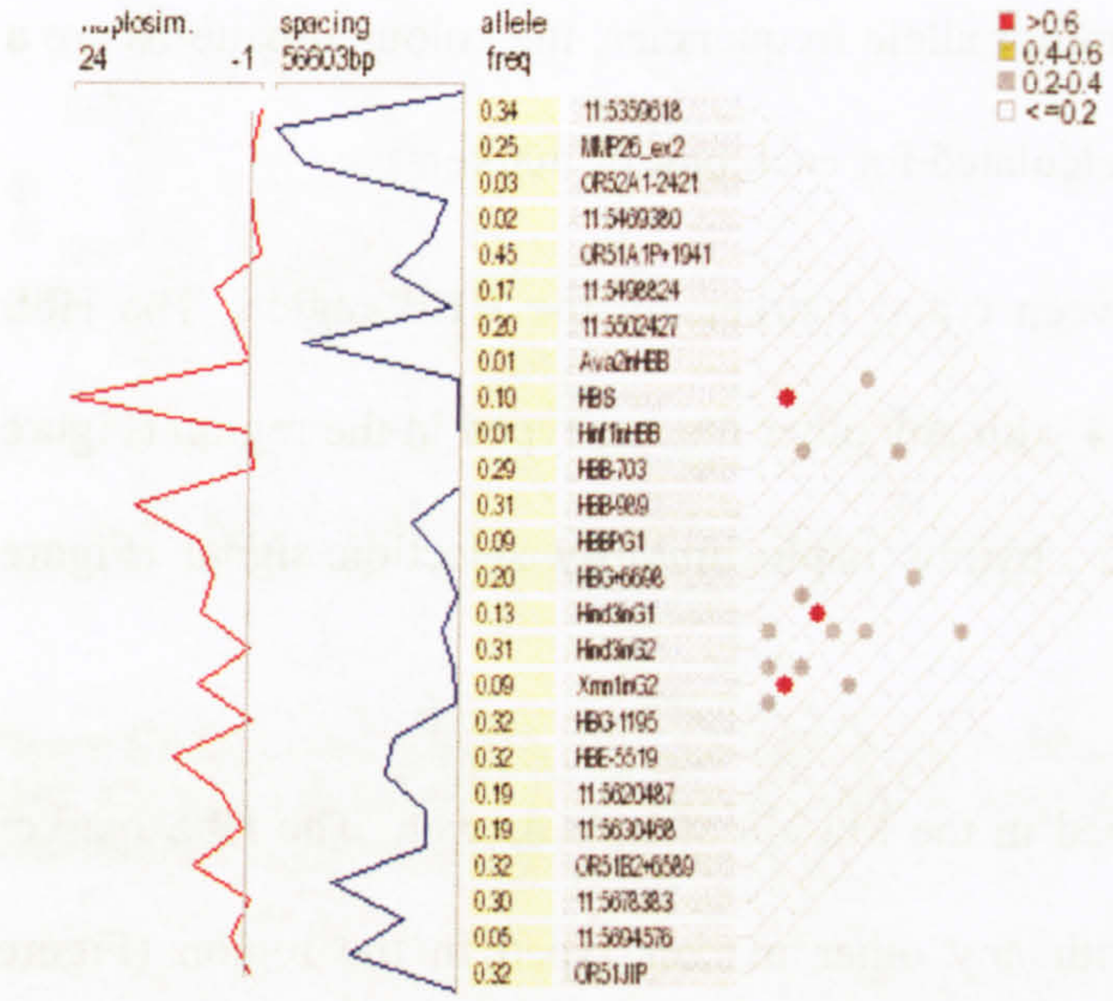
The program **MARKER** was used to generate an LD map of the HBB region in the Hausa and Masalit. A separate map was constructed for each sample (figure 5.5.3.3 and 5.5.3.4). The vertical axis is the SNPs typed and minor allele frequencies, the coloured patterns are a statistical representation of the r^2 value calculated for each pair of markers.

In the Hausa, little LD was found between typed markers in the HBB region. The HbS marker did not have an r^2 value above 0.4 with any other marker typed in the region (Figure 5.5.3.3a). Nevertheless, it displayed an obvious haplosimilarity selection signal (Figure 5.5.3.3b).

The LD was weak between markers typed in the Masalit sample as well. The HbS marker did not have an r^2 value above 0.2 with any other marker typed in the region (Figure 5.5.3.4a).

The selection signal at the HbS locus was smaller than that observed in the Hausa (Figure 5.5.3.4b).

A



B

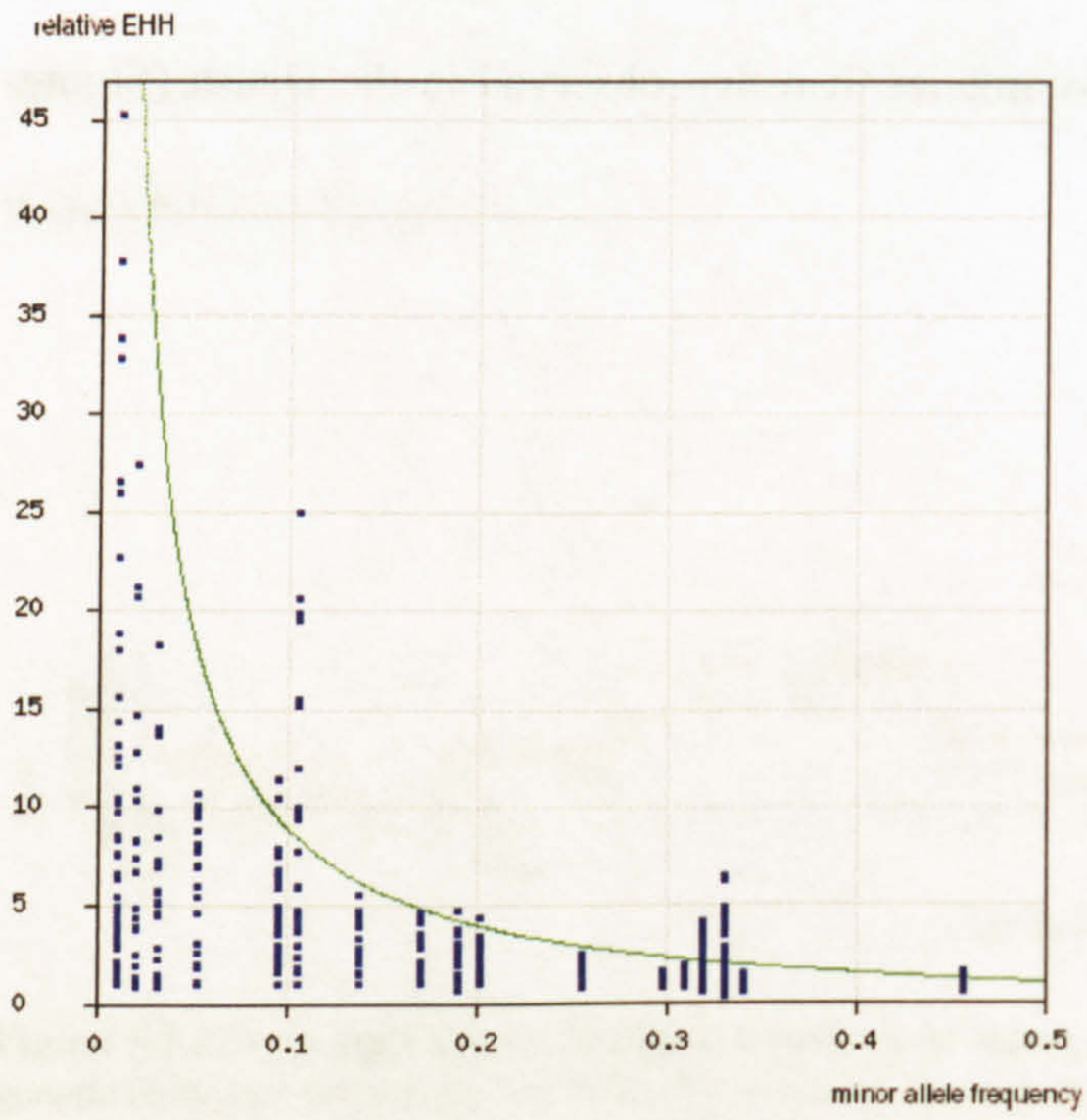
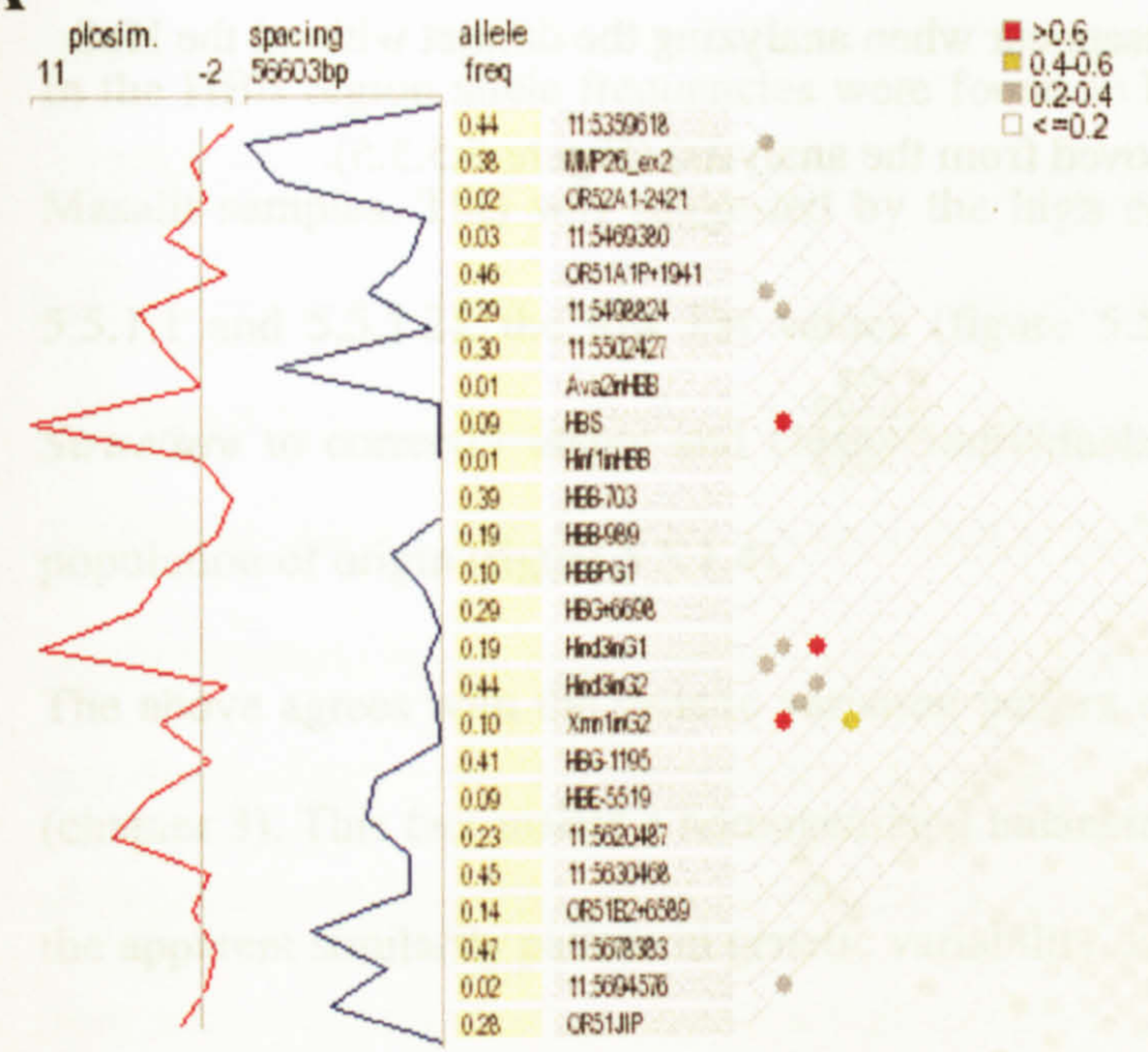


Figure 5.5.3.3: a) Marker Map illustrating the LD between SNPs in the HBB region in the Hausa. LD is measured by r^2 (http://www.gmap.net/marker). Coloured spots connecting SNPs illustrate the LD level between those SNPs. Colour coding is presented in the top right-hand corner. **b) Scatter plot of minor allele frequencies of markers typed in the HBB and their haplosimilarity scores.**

A



B

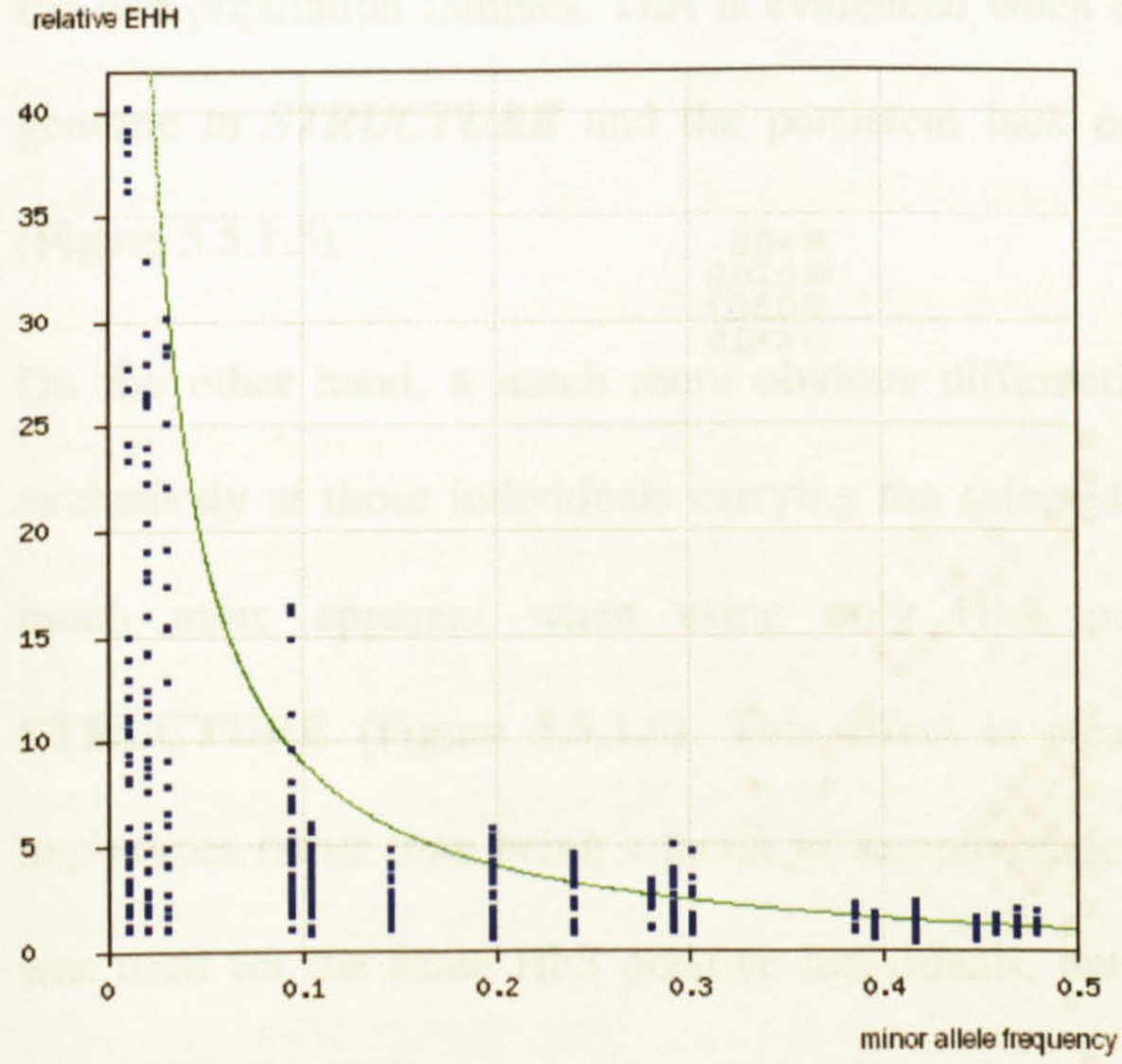


Figure 5.5.3.4: Marker Map illustrating the LD between SNPs in the HBB region in the Masalit sample. LD is measured by r^2 (http://www.gmap.net/marker). Coloured spots connecting SNPs illustrate the LD level between those SNPs. Colour coding is presented in the top right-hand corner. **b) Scatter plot of minor allele frequencies of markers typed in the HBB and their haplosimilarity scores.**

The absence of any other marker with strong LD with the HbS in the 400kb region analysed, meant that the selection signal would disappear when analyzing the dataset without the HbS marker (when the HbS marker was removed from the analysis) (Figure 5.5.3.5).

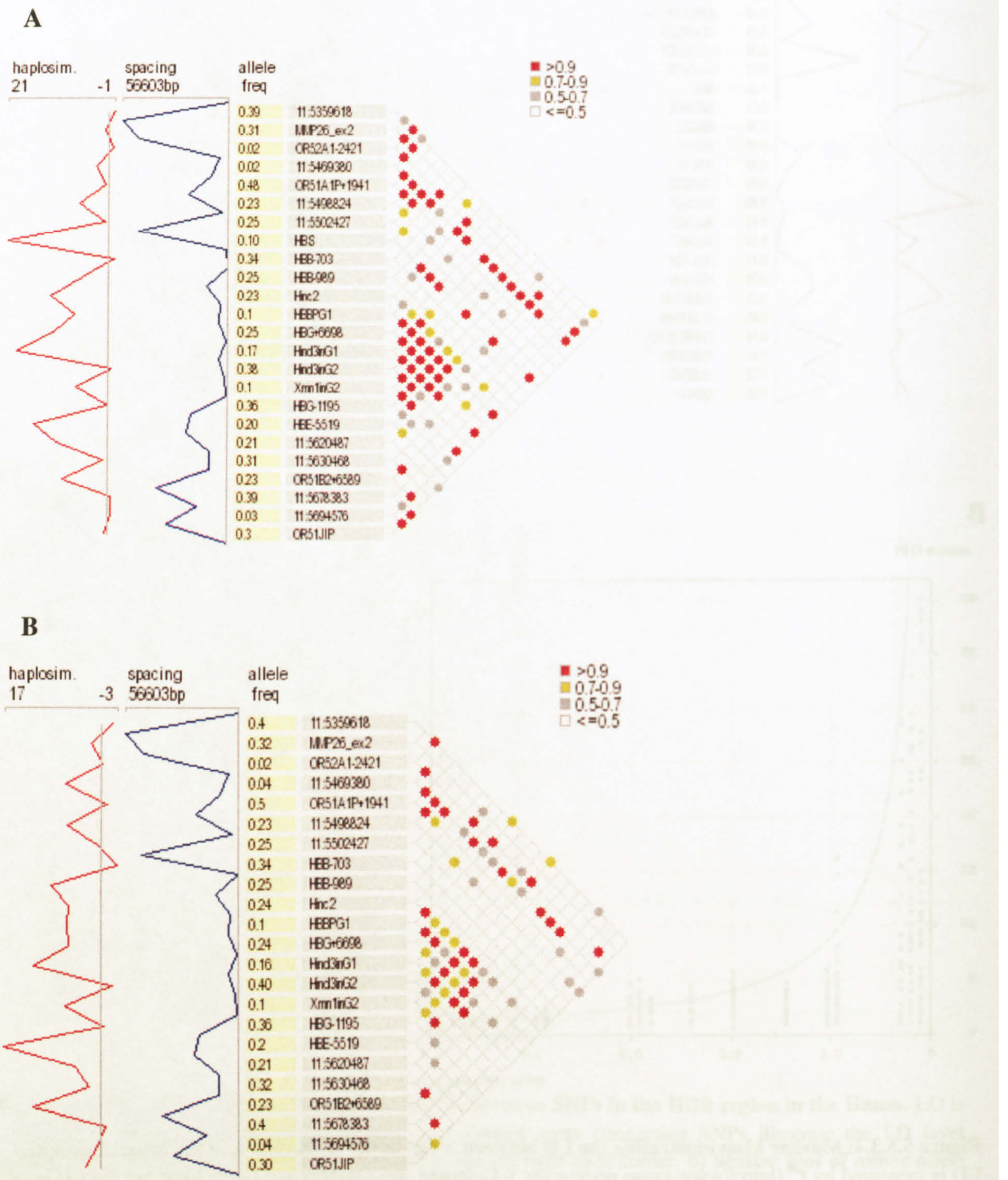


Figure 5.5.3.5: Haplosimilarity scores of markers typed in the HBB region in the combined Hausa and Masalit samples. A) Selection signals when the HbS marker is included in the analysis. B) Selection signals when HbS marker is excluded from the analysis.

5.6. Discussion

In the HBB region allele frequencies were found to be very similar between the Hausa and Masalit samples. This was suggested by the high correlation of allele frequencies (figure 5.5.1.1 and 5.5.1.2), the low Fst values (figure 5.5.1.3), and the failure of the program ***Structure*** to correctly assign and cluster individuals in the combined sample to a distinct population of origin (figure 5.5.1.4).

The above agrees with the genetic variation pattern observed previously in the 5q31 region (chapter 3). This fact makes a homogenizing balancing selection force being responsible for the apparent similarity pattern in genetic variability data a less likely possibility.

In addition to the possibility of the inadequacy of marker resolution, it is likely that the above also reflects the general genetic similarities and low genetic differentiation between the two population samples. This is evidenced when using data from 91 markers across the genome in ***STRUCTURE*** and the persistent lack of sample structuring by the program (Figure 5.5.1.5).

On the other hand, a much more obvious differentiation pattern emerged when looking exclusively at those individuals carrying the selected HbS haplotypes. Structuring became much more apparent when using only HbS positive individuals in the program **STRUCTURE** (Figure 5.5.1.6). This effect is clearly related to selection on the HbS haplotypes rather than being a factor of sample choice. When the same number of markers was used for the same HbS positive individuals, but they were markers at least a 100 kb away from the HbS variant; the pattern of sample structuring disappeared (Figure 5.5.1.7).

The above point suggests that even in cases where the functional variant is the same in different population groups, its different haplotypic background between the groups might be accentuated more so than those of neutral markers. If a genetic variant is under natural

positive selection in several populations, it is likely to reside on different haplotypic backgrounds and thus have a more discrepant LD relationship with other markers than might be the case for other markers. This could be attributed to the faster haplotypic divergence under selection than neutral genetic drift.

Further in-depth analysis of the HbS haplotypes by trying to interpret them in terms of the classically described HbS haplotypes, revealed that in spite of the shared haplotype sequences between the two population groups, the relative frequency of each of these haplotypes is different between these two groups. This further supports the point made above about group differences in selected haplotypes even when it is the same polymorphism that is under selection in the different groups.

The fact that the functional variant (HbS in this case) was found on different haplotypic background in different populations, could complicate results of association studies carried out in different African populations, or studies that suffer from population sub-structuring of the sample. Even though the same variant might be under selection in the compared populations, its haplotypic background is likely to be different.

It is likely for the type of variants targeted in association studies, those functional variants affecting susceptibility, to be acted on by natural selection. My data shows that these selected variants are more likely to have a discrepant LD relationships between different population groups because they could occur on different haplotypic backgrounds, either due to the mutation arising several times in history or due to demography of the populations studied. The haplotype on which the functional polymorphism might have originally occurred could change with time, by recombining and acquiring variations from new mutations, a subset (migration or bottleneck events) of these new haplotypes, might be

introduced into another population and then start to rise in frequency by a selective pressure encountered in the new environment.

If the functional SNP is not typed in the association study, then its detection depends on linkage with nearby markers, and if this relationship is discrepant between groups of different ethnicities considered jointly in the study, then different SNPs could tag the functional one across populations. This could lead to a decrease in the power of any of these tagging SNPs if the sample suffers from undetected sub-structuring, possibly resulting in a few markers clustering in this genomic region displaying association signals below the level of significance. Therefore it might prove a sensible approach in the interpretation of results from association studies to consider a region-wide significance.

Taking a closer look at the haplotype structure in the HBB region in general, and the HbS haplotypes in particular; it becomes apparent that another aspect setting the HbS haplotypes apart from others in the region, in addition to their greater inter-group differentiation, is the fact that the HbS haplotypes are the highest in frequency across the region when compared with the background haplotype structure (Figure 5.5.2.1). This relatively high frequency was maintained as far as 1.7 Mb span around the HbS variant, long after the frequency of all other haplotypes decreased to a single copy. Furthermore, the HbS haplotypes appear in general to be more homogenous when compared with the HbA haplotypes (Figure 5.5.2.2).

Marker spacing and frequency profoundly influence observed levels of haplotype diversity. The number of haplotypes also depends on the number of SNPs and the recombination rate in the genomic area. In the combined Hausa and Masalit sample, the high frequency of HbS carrying haplotypes are in contrast to the high haplotype diversity in the region as a whole, reflecting the high recombination rate. This high frequency was found to extend to a 1.7 Mb region centered on the HbS is a phenomenon which could indicate the positive selection

pressure on the HbS variant sweeping its haplotype to a high frequency over an extensive genetic distance without the chance for recombination to break it down.

It would be interesting to further explore the relationship between high frequency haplotypes and recombination, and if there are other instances in the genome similar to the observed long range high frequency haplotype in the HBB region. If so, how common and what is the likelihood of chance giving rise to a similar picture. This observation has a possible utility as a metric to look for positive selection signals across the genome.

The above observation could be considered the primer for further investigation of this phenomenon in the search for positive selection signals across the genome. Indeed the next two chapters will be on the exploration and application of this phenomenon in data from the HapMap and MalariaGEN whole genome case control study.

Because the HapMap cell lines are publicly available, I will be able to integrate my experimental data in the β-globin region with the genome-wide SNP data to gain new insights into the relationship between classical β-globin haplotypes and SNP variation.

Although there was evidently a clear selection signal at the HbS variant when using the program *Sweep* on the combined population sample (Figures 5.5.3.1 and 5.5.3.2), when examining this signal for individual populations separately, it was not as prominent in the Masalit when compared to the Hausa (Figures 5.5.3.3 and 5.5.3.4). Five of the six high frequency haplotypes in the data belonged to the Hausa group. Although the HbS haplotypes in the Masalit samples displayed some degree of homogeniety compared to the HbA haplotypes as observed in figure 5.5.1.6 and figure 5.5.2.2, no selection signal was detected in the sample. This might indicate that identical high frequency extended haplotypes might be much easier to be detected by metrics of positive selection than the effect of homogeneity of the haplotypic background. The HbS allele in Masalit appears to have been more recently

introduced to this population from several different sources, as revealed by its more diverse haplotypic background.

5.7. Conclusion

The analysis in this chapter provided insights on the selected haplotypes in the Hausa and Masalit. In spite of the fact that HbS was the same selected variant in the two populations, it resided on different selected haplotypes. This suggests that, in association studies, the causal variant may be tagged by different SNPs in different populations, which may lead to a decrease in the power of these tagSNPs to detect disease association if the sample suffers from undetected sub-structuring.

In this chapter I found evidence of positive selection in the β-globin region in the Hausa population. Using known metrics to detect this signal depended on inclusion of the HbS functional variant in the analysis. However, I noted that the high frequency HbS-carrying haplotype extends to a very long distance (1.7 Mb), spanning several recombination hotspots. This raised the question of whether a method might be developed to search for such extended high frequency haplotypes even if the causal variant was not genotyped, and whether this might provide a useful tool for screening for signals of selection in the whole genome. In the next chapter, I go on to examine this question by using the publicly available genome-wide genotyping data of the HapMap project.

Chapter 6:

Genome-wide search for natural selection signals by characterizing extended-high-frequency haplotypes in the HapMap data.

6.1. Abstract

In the previous chapter, analyzing the polymorphism data of the Hausa and Masalit HBB region, I observed an unusually long and high-frequency haplotype that carried the HbS allele. This haplotype was easily identifiable even when the HbS genotype was omitted from the analysis.

Because such phenomenon could be a surrogate for positive selection signals in the genome, I set out to better characterize it by attempting to answer the following two questions: Firstly, is the high frequency extended HbS haplotype exclusive to the Sudanese populations or can it be discernable in another African population, secondly, are there similar instances in other genomic regions, and if so, is there any supporting evidence of them being candidates of positive selection.

I used publicly available genome-wide genotyping data from the Yoruba (YRI) samples that were typed in the HapMap project. To make a meaningful comparison with the Sudanese data, I typed the 90 YRI samples for SNPs typed in the Sudanese samples that had not been typed in the HapMap project. I developed programming scripts that dealt with the large volumes of data generated by HapMap and tested for identical long rage high frequency haplotypes employing a sliding window approach. This involved taking account of the

variation in recombination rate to enable comparisons to be made between different regions of the genome.

In the Yoruba, I identified an HbS haplotype of a strikingly different frequency from others in the region. This haplotype extended to 1.2 cM and was clearly unusual when compared to other haplotypes across chromosome 11. I also identified a few other regions in the genome where similar instances of extended high frequency haplotypes were present, and which had some suggestive evidence of being under selection.

6.2. Objectives

- Investigate whether the phenomenon of a long-range high-frequency HbS haplotype observed previously in the Sudanese population samples is replicated in the HapMap YRI sample.
- Quantify it in the context of the whole genome and identify other similar instances.
- Look at regions identified as outliers and determine whether there is any suggestive evidence of selective pressure.
- Attempt better localization of the functional variant in a region with an unusually extended high frequency haplotype.

6.3. Introduction

Advantage of using a genome-wide empirical approach.

A genome-wide empirical approach might be a sensible way of detecting positive selection signals. This is an approach that does not depend on assumptions about population history, as opposed to the expectations of population genetic models which depend on assumptions about demographic parameters for which estimates remain ambiguous.

Some demographic factors like population growth, subdivision, bottlenecks and admixture can cause departures from the neutral model that are indistinguishable from those caused by natural selection. It is also possible for a departure from the neutral model at any specific locus to be caused by a combination of both population history and selection. One way to overcome this problem is to recognize that population demographic history affects patterns of variation at all loci in a genome, whereas natural selection acts upon specific loci (Przeworski, Hudson et al. 2000; Andolfatto 2001; Nielsen 2001). Thus, by sampling a large number of unlinked loci throughout the genome, it is in principle possible to distinguish between selection and demography.

Available methods for detecting positive selection.

Many of the common population genetic methods for detecting selection are based on comparing variation within and between species, most famously the HKA test (Hudson, Kreitman et al. 1987). In this test, the rate of polymorphisms to divergence is compared for multiple genes. If the ratio varies more among genes than expected on a neutral model, neutrality is rejected.

When a locus shows extraordinary levels of genetic population differentiation, compared with other loci, this may then be interpreted as evidence for positive selection. One of the first neutrality tests proposed, the Lewontin-Krakauer (Lewontin and Krakauer 1973) test, takes advantage of this fact. This test rejects the neutral model for a locus if the level of genetic differentiation among populations is larger than predicted by a specific neutral model. Akey et al. (Akey, Zhang et al. 2002) looked at variation in F_{ST} among human populations genome-wide.

Selection also affects the distribution of alleles within populations. Some of the most commonly applied tests are based on summarizing information regarding the frequency spectrum. Selection against deleterious mutations will increase the fraction of mutations segregating at low frequencies in the sample. A selective sweep has roughly the same effect on the frequency spectrum (Braverman, Hudson et al. 1995). Conversely, positive selection will tend to increase the frequency in a sample of mutations segregating at high frequencies. The most famous example is the Tajima's D test (Tajima 1989). In this test, the average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites (SNPs). If the difference between these two measures of variability is larger than what is expected on the standard neutral model, this model is rejected. Fu & Li (Fu and Li 1993) extended this test to take information regarding the polarity of the information into account by the use of an evolutionary outgroup (e.g., a chimpanzee in the analysis of human genetic variation), and more refinements were introduced by Fu (Fu 1997). Fay & Wu (Fay and Wu 2000) suggested a test that weights information from high-frequency derived mutations higher.

An incomplete sweep (when the adaptive mutation has not yet been fixed in the population) leaves a distinct pattern in the haplotype structure. This has led to the development of many statistical methods for detecting selection based on LD. Hudson et al. (Hudson, Bailey et al. 1994) developed a test based on the number of alleles occurring in a sample. Andolfatto et al. (Andolfatto, Wall et al. 1999) developed a related test to determine whether any subset of consecutive variable sites contains fewer haplotypes than expected under a neutral model. A similar test was also proposed by Depaulis & Veuille (Depaulis and Veuille 1998). A variation on this theme was proposed by Sabeti et al. (Sabeti, Reich et al. 2002) who considered the increase in the number of distinct haplotypes away from the location of a putative selective sweep. The presence of long high frequency haplotypes across the genome was taken as a possible evidence of natural selection (HapMap 2005; Frazer, Ballinger et al. 2007; Sabeti, Varilly et al. 2007).

Finally, the MacDonald-Kreitman test (McDonald and Kreitman 1991) explores the fact that mutations in coding regions are of two types: nonsynonymous mutations and synonymous mutations. It summarizes the data in what has become known as a MacDonald-Kreitman table, which contains counts of the number of nonsynonymous and synonymous mutations within and between species. If selection only affects the nonsynonymous mutations, negative selection will reduce the number of nonsynonymous mutations and positive selection will increase the number of nonsynonymous mutations, relative to the number of synonymous mutations.

Evidence that positive selection is widespread across the genome.

Ambiguity in the interpretation of classical population genetic neutrality tests, due to the presence of confounding demographic factors, may have precluded the establishment of firm conclusions regarding the pervasiveness of selection. As more large-scale data have

accumulated, and methods that are robust to demographic assumptions have been applied, a clearer picture of the pervasiveness of positive selection has been established.

Several scans of the human genome have been undertaken to search for regions under natural selection, and more are underway (Cargill, Altshuler et al. 1999; Sunyaev, Lathe et al. 2000; Stephens, Schneider et al. 2001; Akey, Zhang et al. 2002; Payseur, Cutter et al. 2002).

There is an increasing amount of evidence that selection is important in shaping variation within and between species. In human SNP data, there is a clear difference in the frequency spectrum between non-synonymous and synonymous mutations (Williamson, Hernandez et al. 2005). This observation shows that a large proportion of the mutations that are segregating in humans are affected by selection. In addition, there is a rapidly growing list of specific genes that show evidence for positive selection in both humans and other organisms (Bamshad and Wooding 2003; Vallender and Lahn 2004). This explosion of results showing a presence of positive selection may suggest that positive selection is much more common than previously believed. Standing levels of variation in the genome can be explained by the proposed models of repeated selective sweeps (Gillespie 2000). In these models, known as genetic draft models, mutations causing species differences are not neutral mutations increasing in frequency due to genetic drift, but primarily neutral mutations increasing in frequency due to linkage with adaptive mutations sweeping through the population. Even though only few mutations are adaptive, the population genetic dynamics is determined by the selective forces acting on the adaptive mutations, not by genetic drift, and as yet there is no mathematical or empirical evidence to suggest that this model is unrealistic.

Limitations of available methods.

The power of tests used for detecting natural selection is typically determined by carrying out simulations under a restricted range of demographic models and parameters to estimate the critical values that support rejection of the neutral model (Simonsen, Churchill et al. 1995; Fu 1997). To this end, an understanding of population history is crucial for identifying the genes that are subject to selection.

The neutrality tests are all tests of complicated population genetic models that make specific assumptions about the demography of the populations, in particular a constant population size and no population structure. In addition, in some of the tests there may be other implicit assumptions regarding distributions of recombination rates and mutation rates. Many of these tests have long been known to be highly sensitive to the demographic assumptions. For example, Tajima's D test (Tajima 1989) would reject a neutral model very frequently in the presence of population growth (Simonsen, Churchill et al. 1995). Simple models of population subdivision can lead the commonly used neutrality tests to reject the neutral model with high probability, even in the absence of selection. In addition, even if the presence of selection can be established, in many cases it can be difficult to distinguish between the pattern left by selective sweeps and selection on slightly deleterious mutations (background selection) (Charlesworth, Morgan et al. 1993).

Because of the effect of demographic assumptions on the population genetic neutrality tests, the results of these tests have often been contentious and often have not led to firm conclusions regarding the action of selection. It is not very meaningful to reject the standard neutral model using these methods without paying careful attention to the underlying demographics.

One possible way to circumvent the problem of demographic confounding effects is to compare multiple loci. The assumption being that if strong departures from the neutral model are seen only on one or a few outlier loci, this may be interpreted as evidence for selection on these loci.

Standard methods for detecting selection from population genetics can, in principle, be applied to provide a detailed picture of the regions of the genome that may have been targeted by selection. However, most SNP data have been obtained through a complicated SNP discovery process that involves the discovery (or ascertainment) of SNPs in a small sample followed by genotyping in a larger sample, or by choosing high frequency markers from publicly available databases. The process by which the SNPs have been selected affects levels of LD observed in the data and the frequency spectrum, which makes these studies not ideally suited for detecting selection, because ascertainment bias complicates downstream analyses, one example of that might be the skewness it creates in allele frequency spectrum (Akey, Zhang et al. 2003).

Current methods for detecting selective sweeps have little or no robustness to the demographic assumptions and varying recombination rates, and provide no method for correcting for ascertainment biases, as well as the limitation of requiring DNA sequence data by many of these tests.

What is known about the extent of the signals of positive selection.

Apart from very few recent studies, very little was done in the past in way of establishing how far the effects of positive selection extend in the human genome. The feasibility of such analysis nowadays could be due in part to the recent improvement in efficiency of high throughput genotyping technologies.

How far a selective sweep extends depends on the combined effects of selective pressure strength, time since it occurred, background recombination rate and population demographic history.

A few recently published studies indicate that selection can lead to non-random associations among SNPs over great physical and genetic distances, in one study (Saunders, Slatkin et al. 2005) a genomic region of approximately 1.6 Mb around G6PD was characterized by long-range LD. In another study (Yu, Sabeti et al. 2005) a region of approximately one megabase of human chromosome 12 shows extensive LD, this effect was subsequently attributed to selection of a pre-expansion CAG repeat within exon 1 of the Spinocerebellar ataxia type 2 gene (SCA2).

Purpose of this investigation.

Here I will explore a genome-wide approach that uses empirical distribution of data to eliminate variability created by demographic factors. This may provide a rapid way to define the area in the genome over which it would be useful to carry out further, more refined analysis to look for the adaptively important functional variant. This initial definition of area is the most important aspect of this analysis that would aid downstream analysis that would have otherwise yielded false negatives if used over too small or too large a genomic area.

Although this method requires both haplotypic and recombination information, it is less computationally intensive than other methods based on the extended haplotype homozygosity (Sabeti, Reich et al. 2002) which in addition require partitioning at each SNP or group of SNPs and then comparing the two partition groups in terms of how similar the haplotypes are in each group. Furthermore, this method circumvents the problem of SNP ascertainment bias because it looks at such a large genomic area, the choice of SNPs plays little or no role in haplotypes' determination, which makes it greatly robust in picking up

regions where positive selection has played a role even if the functional variant or a tightly linked marker has not been typed, rather it makes use of the more stable information contained in the long range haplotype carrying the selected mutation without the need to actually type it.

6.4. Materials and Methods

6.4.1. Subjects and DNA preparation

Ninety DNA samples used in phase1 HapMap project from the Yoruba in Ibadan, Nigeria http://www.hapmap.org/abouthapmap.html .

Publicly available HapMap data was also used for thirty U.S. trios provided samples, which were collected in 1980 from U.S. residents with Northern and Western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH). The blood samples were converted into cell lines, which are used to make DNA, by the non-profit Coriell Institute for Medical Research http://locus.umdnj.edu/nigms. Cell lines of the relevant samples were ordered from the Coriell Cell Repository at Coreill Institute for Medical Research (see <http://locus.umdnj.edu/nigms/>) as transformed B-lymphocytes from peripheral blood, cell lines were cultured in the lab and then DNA was extracted using CST Genomic DNA Purification Kit (<http://www.chargeswitch.com/>). DNA was subsequently quantified using picogreen and NanoDrop technology <http://www.nanodrop.com/>, concentration standardized to 20 ng/μL, whole genome amplification using Primer Extension Pre-amplification PEP (Zhang, Cui et al. 1992) was carried out on these samples in a 50 μL reactions. PEP is a method which uses random primers 15 base pairs long to amplify the

whole genome by PCR. The products are usually in the region of 1.5 kb in length. Samples were tested with an ARMS reaction to determine quality of DNA before genotyping.

6.4.2. Genotyping

Sixteen markers in the HBB region were typed in the 30 HapMap YRI trios (Table 6.4.2). Markers were chosen to overlap with those previously typed in the Sudanese samples, excluding those for which data is already available in phase 1 HapMap dataset. Those markers chosen included the five RFLP marker described previously in chapter 5. The RFLP markers were chosen from literature to define the previously described classical β-globin haplotypes (for details see Materials and Methods chapter and chapter 5). Other than the RFLP markers, genotyping was carried out using primer-extension / mass-spectrometry (Sequenom) technology.

MARKER ID	SNP ASSAY NAME	RS NUMBER	POSITION
1	rs7114854	rs7114854	11:5100498
2	11:5498824	rs4910722	11:5153293
3	11:5502427	rs4910726	11:5156896
4	11-5519408	rs4910732	11:5173877
5	HBS	rs334	11:5204808
6	hHbCB	rs33930165	11:5204809
7	Hinf1 in HBB	rs10742584	11:5205346
8	HBB-703	rs11036364	11:5205580
9	HBB-989	rs16911905	11:5205866
10	HincII	rs968857	11:5217034
11	Hind3 in G1	rs6578593	11:5226375
12	Hind3 in G2	rs2070972	11:5231293
13	Xmn1 in G2	rs7482144	11:5232745
14	rs7938837	rs7938837	11:5319683
15	rs7929631	rs7929631	11:5324552
16	rs1498468	rs1498468	11:5367607

Table 6.4.2: Assays of SNPs typed in the β-globin region in the HapMap YRI sample.
Listed are the Laboratory assay names, rs reference numbers as well as the location of SNPs on chromosome 11 (Ensembl release 39).

6.4.3. Bioinformatics and statistical analysis

I downloaded haplotypes and recombination rate estimates for the phase 1 HapMap Yoruba and CEU population from the HapMap site. For these datasets **PHASE (v2.1)** was used to infer haplotypes for the 1 million–SNP of HapMap release 16.c, and recombination rates were estimated using the coalescent method of McVean et al. (McVean, Myers et al. 2004). The method uses a probabilistic model (the coalescent) to describe patterns of genetic variation in which the genetic map is a parameter that can be estimated from data. The method is implemented within the LDhat package (http://www.stats.ox.ac.uk/~mcvean/LDhat).

Recombination rates were estimated separately for each HapMap analysis panel (YRI, CEU, CHB+JPT). Recombination rates were averaged across populations (www.hapmap.org).

PHASE (v2.1).

I used **PHASE** software package, version 2.1 (http://www.stat.washington.edu/stephens/software.html) (Stephens, Smith et al. 2001) (Stephens and Donnelly 2003) to infer the haplotypic phase from the genotypic data I generated in the laboratory. The reason I chose this software for haplotype inference in my data was that **PHASE (v2.1)** was shown to be the most accurate algorithm in a comparison between methods used for phase inference of haplotypic data from genotypic data (Marchini, Cutler et al. 2006).

MARKER.

MARKER was used to calculate and graphically present LD between markers and to display the Haplosimilarity values in the β-globin region of chromosome 11 for the HapMap YRI (http://www.gmap.net/marker).

Haplosimilarity

The haplosimilarity test implemented in the **MARKER** application (http://www.gmap.net/marker) (Hanchard, Rockett et al. 2006).

Extended high-frequency haplotype analysis

I developed a Perl script (http://www.perl.org/) to run on a UNIX platform in order to look for instances of unusually extended high-frequency haplotypes in the genome. The Perl code was used to scan the human genome employing an overlapping window approach. The script ext_hap_freq.pl (see appendix) looks at all the haplotypes within a predefined window. All chromosomes of both the YRI and CEU were scanned using window size prefixed to 360 markers. The window slides across haplotypes supplied (phased HapMap data) by shifting the window position along the length of each chromosome. Firstly, I used window shift of 180 markers, so as to make windows overlap by half their sizes. Then I carried out another genome scan with the much smaller window shift of 1 marker.

The outputs for each position of the window are the numbers of identical haplotypes, in descending order. A typical line could look like: 7, 3, 2, 2 which means that that window (i.e. window of fixed size but starting at that position) had 7 haplotypes the same (throughout the window), and also 3 haplotypes the same, then 2, then another 2, and the rest were distinct. However if all the haplotypes are distinct, the output is1in a single line, so that this (very common) case is countable. This script also calculates the average recombination rate for each window position using a file of estimated recombination rates downloaded from HapMap.

After acquiring the data for all the windows across each of the chromosomes, the frequency of the highest identical haplotype in a window would be plotted against the genetic distance

value for that particular window. This would create a chromosome-wide distribution amenable to an outlier analysis of all windows across a chromosome.

Using excel sheet macros a regression line with the upper and lower 95% confidence intervals was fitted to the scatter plot of maximum haplotype frequency and genetic distance of each window.

6.5. Results

6.5.1. Long-range high frequency HbS haplotypes in YRI

I combined the data that I generated in the laboratory by genotyping the YRI samples, and data downloaded from the HapMap website for the 400 kb region surrounding the HbS marker position (Chr11:4999517 – Chr11:5401780). This resulted in marker density of about 1 marker every 2.4kb using165 markers from HapMap release 16c.1. Phasing was carried out using the software **PHASE** v2.1. Upon analysis of the resulting haplotypes frequencies, it was noted that the highest haplotype frequency was that of an HbS haplotype (Figure 6.5.1.1).

When compared with the Sudanese sample data, this observation was clearer in the YRI sample probably due to the higher marker resolution (165 markers in YRI as opposed to 26 markers typed in the same area in the Sudanese sample).

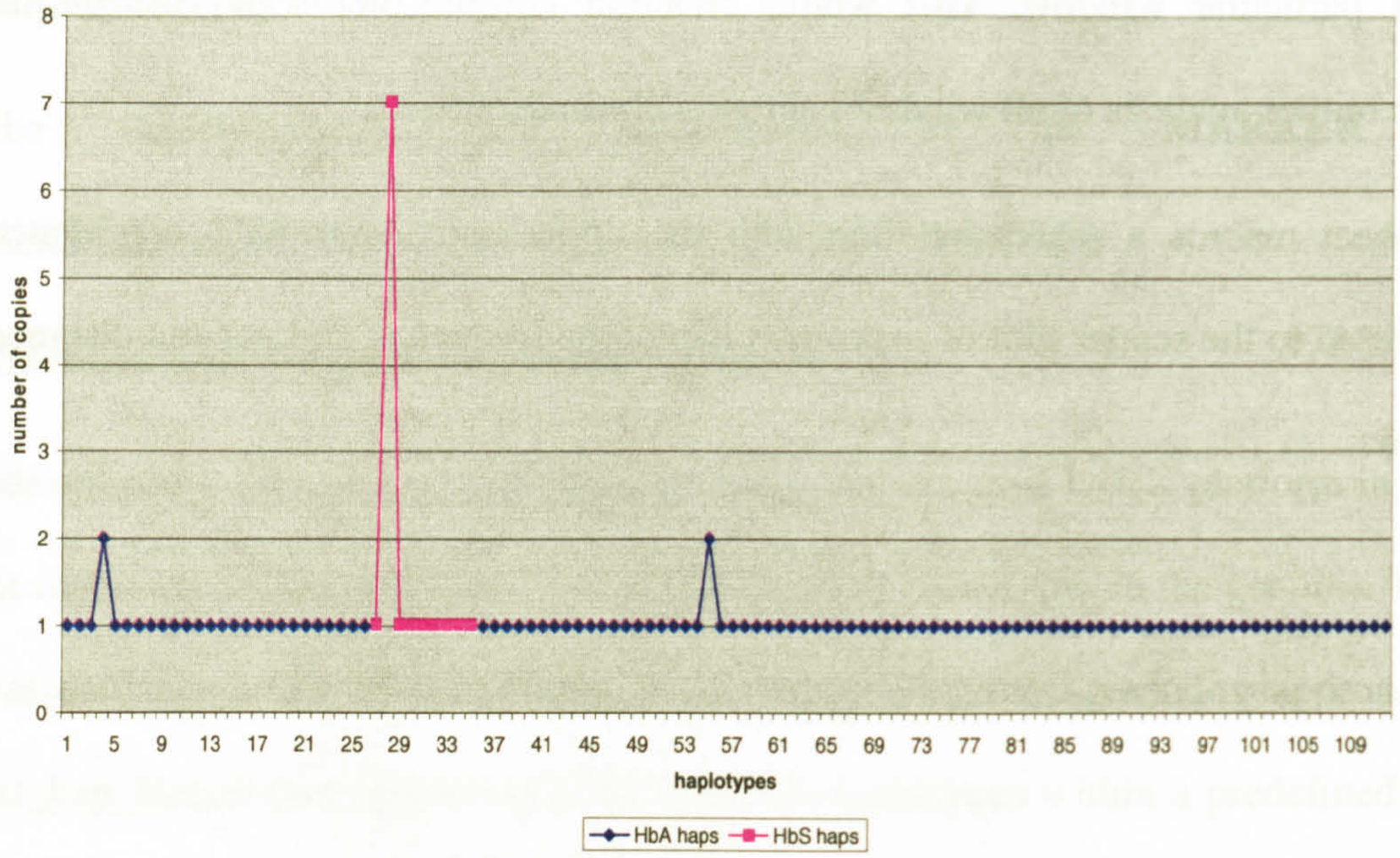


Figure 6.5.1.1: Haplotype frequencies in a 400kb region around HbS in the HapMap Yoruba sample.

On the x axis distinct haplotypes are arrayed by their id numbers. On the y axis the number of identical copies of each haplotype are shown. Blue dots represent HbA haplotypes and pink dots represent HbS haplotypes.

The striking difference between the HbS carrying haplotype and the rest of haplotypes in the same region was clearer in the YRI dataset when a window centred on the HbS was gradually increased in size to include incrementally larger areas around the HbS marker position. The frequencies of identical haplotypes were plotted for each window size (in figure 6.5.1.2 the x axis represent windows of increasing sizes (in kb) centred around the HbS marker while the y axis is the absolute frequencies of identical haplotypes in each window. Red dots represent HbS haplotypes). An HbS haplotype was noted to have high frequency when compared to the frequencies of other haplotypes within the same window. This high frequency was maintained for 1.2 Mb around the HbS allele before declining very rapidly to become indistinguishable from others in the same region. Also noted was the fact that over distances less than 200kb the HbS haplotypes were grouped together with other HbA haplotypes because they were indistinct at the analysed marker density. For a selective sweep of a similar magnitude to the one created by the HbS mutation, the signature of a

single extended high frequency haplotype would be most useful to explore with windows sizes between 200 kb and 1.2 Mb.

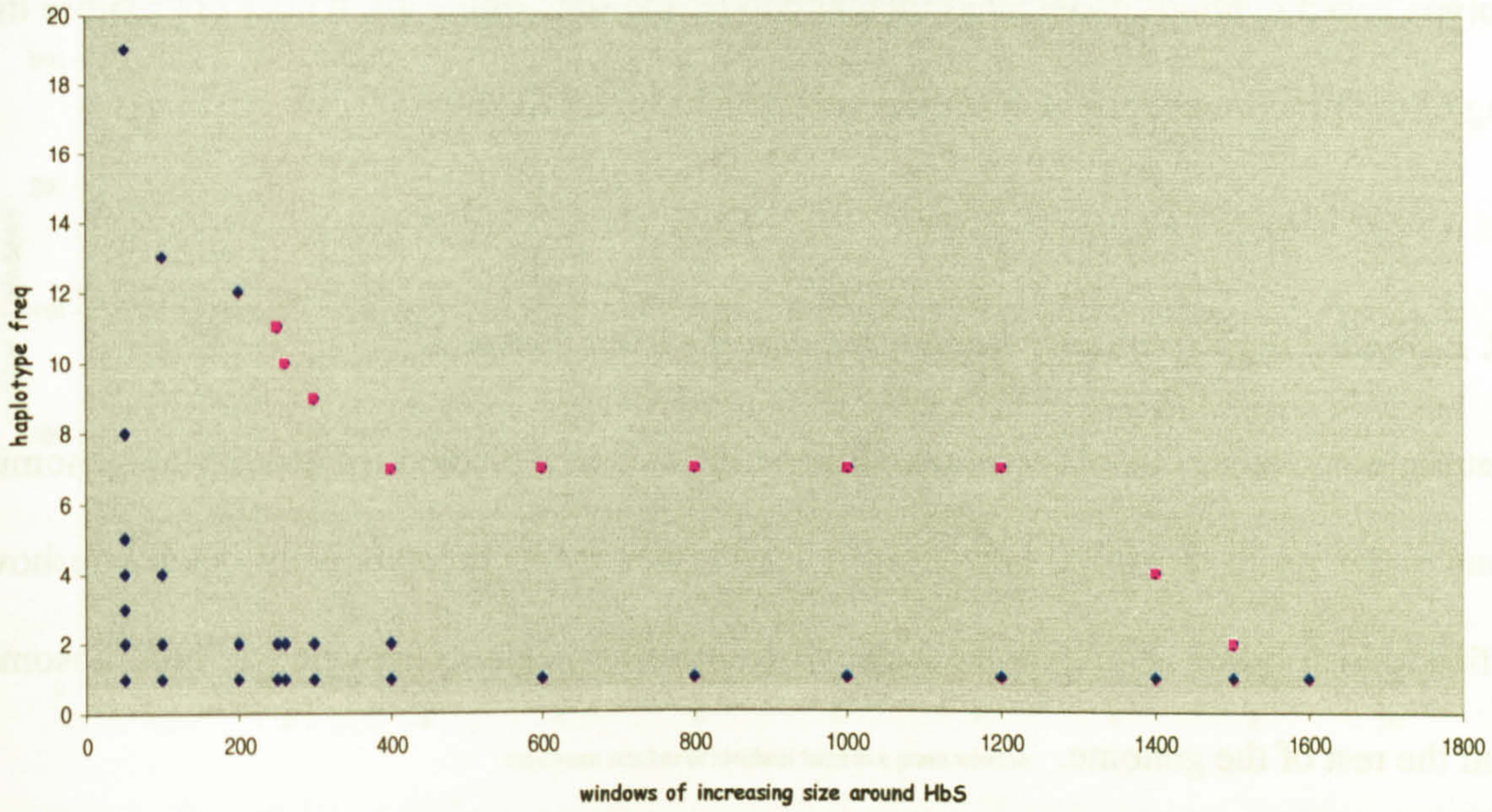


Figure 6.5.1.2: Frequencies of haplotypes in windows centred on the HbS marker and incrementally increased in size. The x axis represent sizes of windows (in kb) centred on the HbS marker. The y axis shows the absolute frequencies of identical haplotypes in each window. Red dots represent HbS haplotypes.

The HbS carrying haplotype maintained its high frequency undisturbed over a 1.2 Mb region (Chr11:4695489- Chr11:5930724). Marker density about 1 marker every 2.6 kb. 430 markers from HapMap release 16c.1 plus the HbS marker), in spite of the presence of several recombination hot spots (Figure 6.5.1.3). Hotspots were defined as areas where there is at least a fivefold increase in estimated local recombination rate (McVean, Myers et al. 2004).

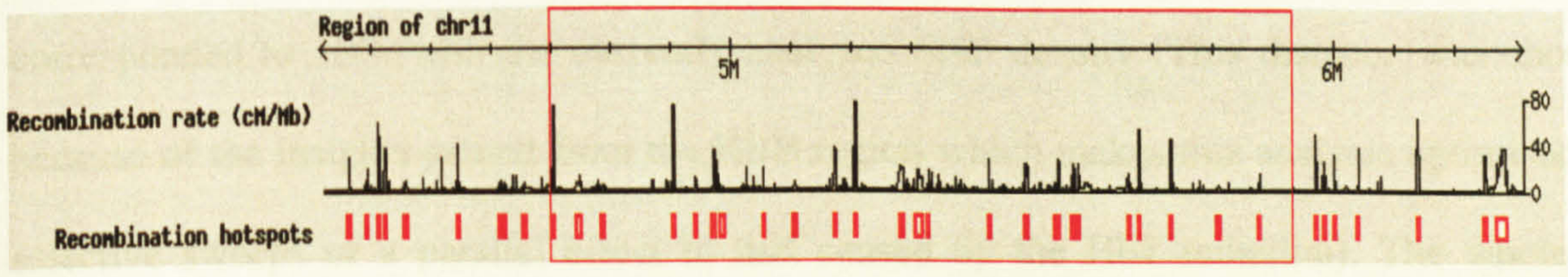


Figure 6.5.1.3: Recombination rate and hotspots in the 1.2Mb region around HbS as estimated from phase 1 HapMap data. Figure downloaded from HapMap website.

The development of an easy method that utilizes the above observations could potentially yield critical clues about regions under natural positive selection because unlike other haplotype based methods considered by many to be the state of the art, it does not require the typing of the functional variant or a marker that is in high LD with it.

6.5.2. Extended high frequency haplotypes across chromosome 11

To determine whether this observation could be utilized as a method for identifying genomic regions under positive natural selection, it was important to quantitatively determine how significant this finding is on a larger scale, when measured against the whole of chromosome 11 and the rest of the genome.

To see whether there are other instances in chromosome 11 where the maximum identical haplotype copies would exceed that of the HbS haplotype over approximately 1.2 Mb distances, windows of 400 markers in size and a 100 marker shift were analysed across chr11. The maximum haplotype frequency in each window was calculated. It was found that other than the 7 identical HbS haplotypes in the HBB region, windows with a maximum haplotype frequency above 6 identical copies were only located in the centromeric region of chromosome 11 where recombination rate is close to zero (Figure 6.5.2.1).

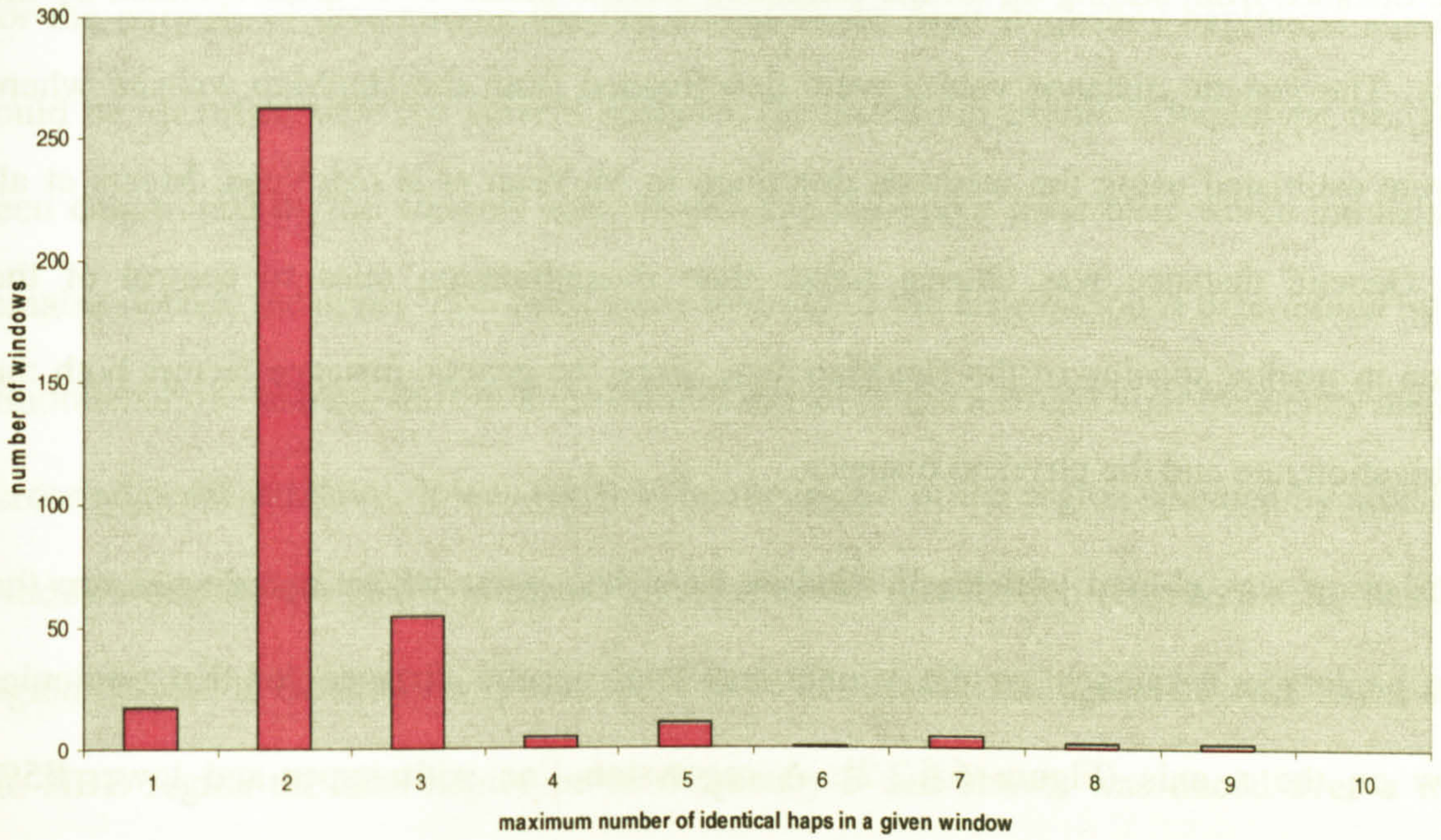


Figure 6.5.2.1: Frequency distribution of maximum number of identical haplotypes in widows across chromosome 11. Windows were of 400 markers in size and shifted by a 100 markers. On the x axis are the bins for the maximum number of identical haplotypes, and on the y axis the number of windows with in each bin.

To distinguish whether the high frequency haplotypes were the result of selection or lack of recombination, it was necessary to account for the recombination rate in genomic areas that were analysed.

Phased haplotypic data from HapMap phase1 was analysed for chromosome 11 using a sliding window approach. A window of a pre-defined number of markers and shift was run across the chromosome. Window size was fixed to 360 markers, because it roughly corresponded to 1Mb with the currently analysed SNP density (This distance was chosen because of the insights gained from the HBB region which makes this analysis optimised to selective sweeps of a parallel effect to that caused by the HbS mutation). The whole of chromosome 11 was analysed using a window shift of 180 markers, and then re-analysed using a 1marker window shift in order to make sure no areas were missed.

In each window, frequencies of all distinct haplotypes were calculated as well as the total genetic distance from adding up all the estimated genetic distances of areas spanned by the window. The genetic distance values were downloaded from the HapMap website where they were estimated using the methods described in McVean et al (McVean, Myers et al. 2004). Genetic distance was chosen rather than recombination rates to control of the variation in marker spacing of the HapMap data, since the genetic distance factors both the recombination rate and the physical distance.

A distribution was plotted with each window as a data point whose coordinates are the highest haplotype frequency on the y axis and total genetic distance for that particular window on the x axis (Figure 6.5.2.2). A regression line with upper and lower 95% confidence interval curves were fitted to the distribution. Any data point outside the upper 95% CI was considered as a candidate for a genomic area undergoing selective sweeps.

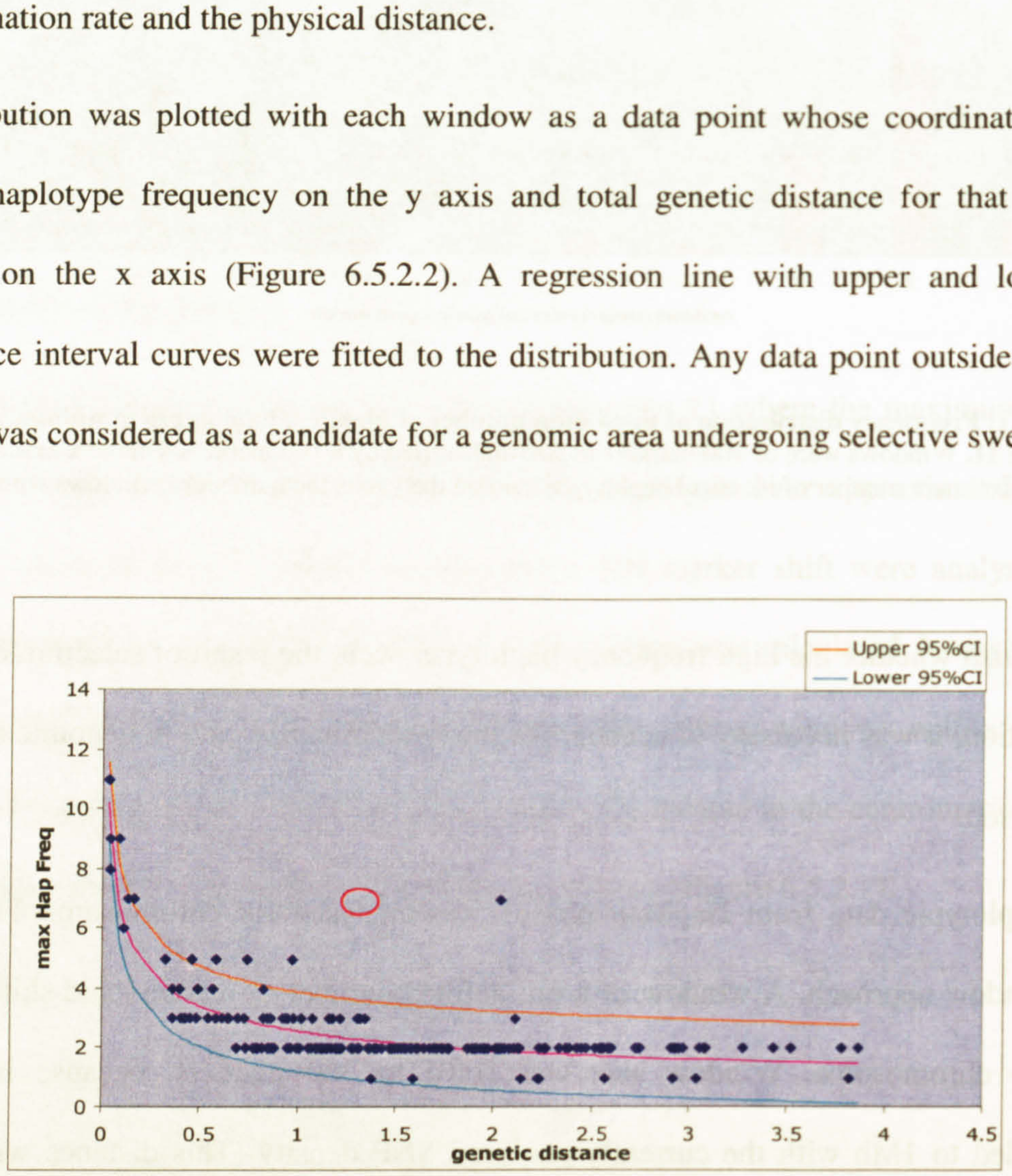


Figure 6.5.2.2: Distribution of highest haplotype frequency and genetic distance values of windows across chromosome 11. Windows were of 360 marker size and 180 marker shift. On the x axis is the genetic distance values in cM and on the y axis the frequency of the haplotype with the most identical copies in a window. Red circle highlights the HBB region.

6.5.3. Another way to determine the extent of the high frequency haplotype

For the purpose of determining the full extent of the high frequency haplotype signal that could be identified with the present analysis, the minimum extent of haplotype has already been determined by the window size chosen, but the upper limit over which the haplotype remains outside the upper 95% confidence interval of the distribution is determined by using windows of one marker shift to trace the full extent of the a single high frequency haplotype across adjacent windows. It was taken to be the length of the region spanned by all adjacent windows outside the 95% confidence interval of the distribution (Figure 6.5.3). The whole region defined in the above way was taken to be of possible biological interest. In the case of the HBB region the HbS haplotype of frequency 5 and above, was traced over a 1.4 Mb region.

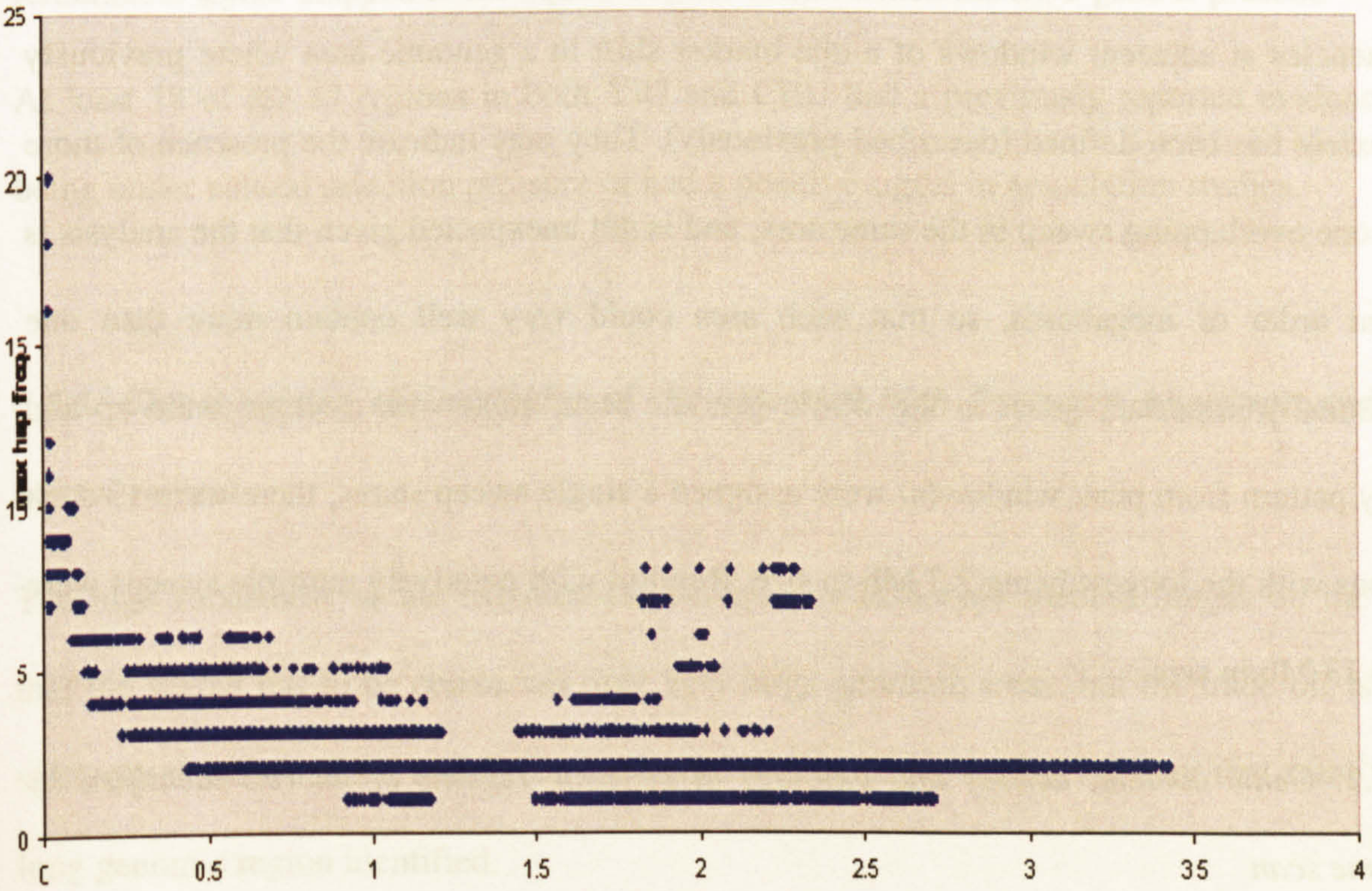


Figure 6.5.3: Distribution of highest haplotype frequency and genetic distance values of windows of 360 marker size and 1 marker shift across chr11. Each data point represents a single window with its genetic distance -in cM- value on the x axis and on the y axis the frequency of the haplotype with the most identical copies in that window.

6.5.4. Regions of interest in HapMap YRI and CEU genomes

Phased haplotypic data from HapMap phase1 for both the YRI and CEU populations were analysed for each chromosome at a time, using the same sliding window approach described previously for chromosome 11 in YRI. In total there were 55 regions that were picked up from the analysis. 23 in YRI and 32 in CEU. From the total 55 regions there were 8 regions shared between YRI & CEU and 39 regions exclusive to one or the other population. The average size of region was (2.78 Mb) in YRI and (2.64 Mb) in CEU samples. The combined length of these regions across the whole genome in bp was found to be 63.8 Mb in YRI which is equivalent to 2% of the total genome size. In CEU it was 84.6 Mb which represent 3% of total genome size. (See appendix for complete details of regions identified by scan).

Sometimes there will be more than one local maximum for the haplotype frequencies in a genomic region, as evidenced by tracing the build-up and decay pattern of haplotype frequencies at adjacent windows of a one marker shift in a genomic area where previously the extent has been defined (described previously). They may indicate the presence of more than one overlapping sweep in the same area, and is not unexpected given that the analysis is in the order of megabases, so that each area could very well contain more than one functionally important gene. In the whole genome scan, areas with a single build-up and decay pattern from peak window(s) were assigned a single sweep status, there were 19 such regions with the longest being 2.7 Mb in size. Regions with putatively multiple sweeps were up to 13 Mb in size.

6.5.4.1. Genic content, density and ontology in genomic regions of interest identified by genome scan

The total number of genes in the regions that stood out from YRI and CEU whole genome scan as possible candidates of positive selection, was 691 genes in a total area of 124002522

bp with average gene density of one gene every 179454 bp. This density was found to be less than average genomic gene density of one gene every 97607 bp when compared with the rest of the genome. (See appendix for a full list of genes in those regions).

Out of these 691 genes identified, 77 genes (10%) were immune genes. When this is compared with the 770 immune genes out of the 33524 total genes in the human genome (about 2%), it becomes clear that there is a higher preponderance of immune genes in the outlier regions identified by the extended-high-frequency-haplotype genomic scan.

6.5.4.2. Supportive evidence from previous studies for scan regions as biologically important

The HBB region had the highest signal in the YRI genome, which is hardly surprising given the fact that the whole scan was optimised on this signal. In the CEU genome the most remarkable signal mapped to the 2q21.3 region within which the LCT gene is present.

At least 18 of the 47 regions in both YRI and CEU had a previously reported evidence of being under natural selection pressure or had a positive signal in association studies.

6.5.5. Characterizing the attributes of the extended high frequency haplotype and the causal variant

The high robustness of the extended-high-frequency-haplotype method hinges on the fact that the search has to be conducted over very large genomic areas, but the trade off is that subsequent more refined search for the causal variant has to be carried over that same very long genomic region identified.

In a region with a selected high frequency extended haplotype, some SNPs might correlate with this haplotype more than others. These SNPs are of most interest for further

investigation because they might either include or closely tag the causal SNP. These SNPs are expected to have an unusually extensive LD pattern compared with other markers in the region. Let us imagine that the minor allele of SNP A is exclusively present on the extended high frequency haplotype. For any other SNP in the region spanned by that haplotype, only one allele will be associated with the minor allele of SNP A because the genotypes at every position of the extended high frequency haplotype are identical. Thus all other SNPs will be in strong LD with SNP A.

In an attempt to identify those SNPs of interest and localize the causal polymorphism in a genomic region spanned by a high frequency selected haplotype, I again used the HBB region in YRI as a model. Equipped by the knowledge of the precise location, frequency and haplotypic and LD relationships of the causal variant (HbS), I looked more closely at an area of 1.1Mb around the HbS (Chr11:4703080- Chr11:5796500. Marker density about 1 marker every 2.6 kb using 416 markers from HapMap release 16c.1 plus the HbS marker).

I carried out several attempts to summarize and compare LD between markers in the pre-defined genomic area, using different LD statistics. I chose the statistic ($|D'| - \Delta^2$) to represent LD instead of the more common $|D'|$ or Δ^2 because the distribution curve was noted to become more uniform with this measure (see below). The reason why subtracting the value of Δ^2 from absolute D prime leads to more uniform distribution is not fully understood, but this kind of manipulation of the data could be justified by the fact that all data points of the empirical distribution underwent the same transformation. Based on this statistic I formulated a metric that I will refer to as the LD-summary statistic (LDSS) to help localize the SNPs of most interest in the HBB region for further analysis. For this purpose a Perl script was used to create an output file with an average LD-summary statistic for each marker in the defined genomic region. Input to the script required a file with genotype data for markers typed in the genomic region of interest.

First, LD-summary statistic was calculated for each marker by firstly calculating the LD between that marker and every other marker in the region using the EM algorithm (LD relationships were described by subtracting the delta^2 (Δ^2) value from the absDprime value). For each marker the sum of all its LD values was divided by the number of markers in the data minus one (the number of relationships) to get the average LD value for this marker. This was done for all the markers in the defined genomic area. The distribution of LDSS and MAF of markers was then plotted. In the resulting scatter plot each marker was represented by a data point whose horizontal position was determined by the value of its MAF and vertical position is determined by the value of its LDSS (table 6.5.5.1) (see appendix for Perl script used to run the analysis).

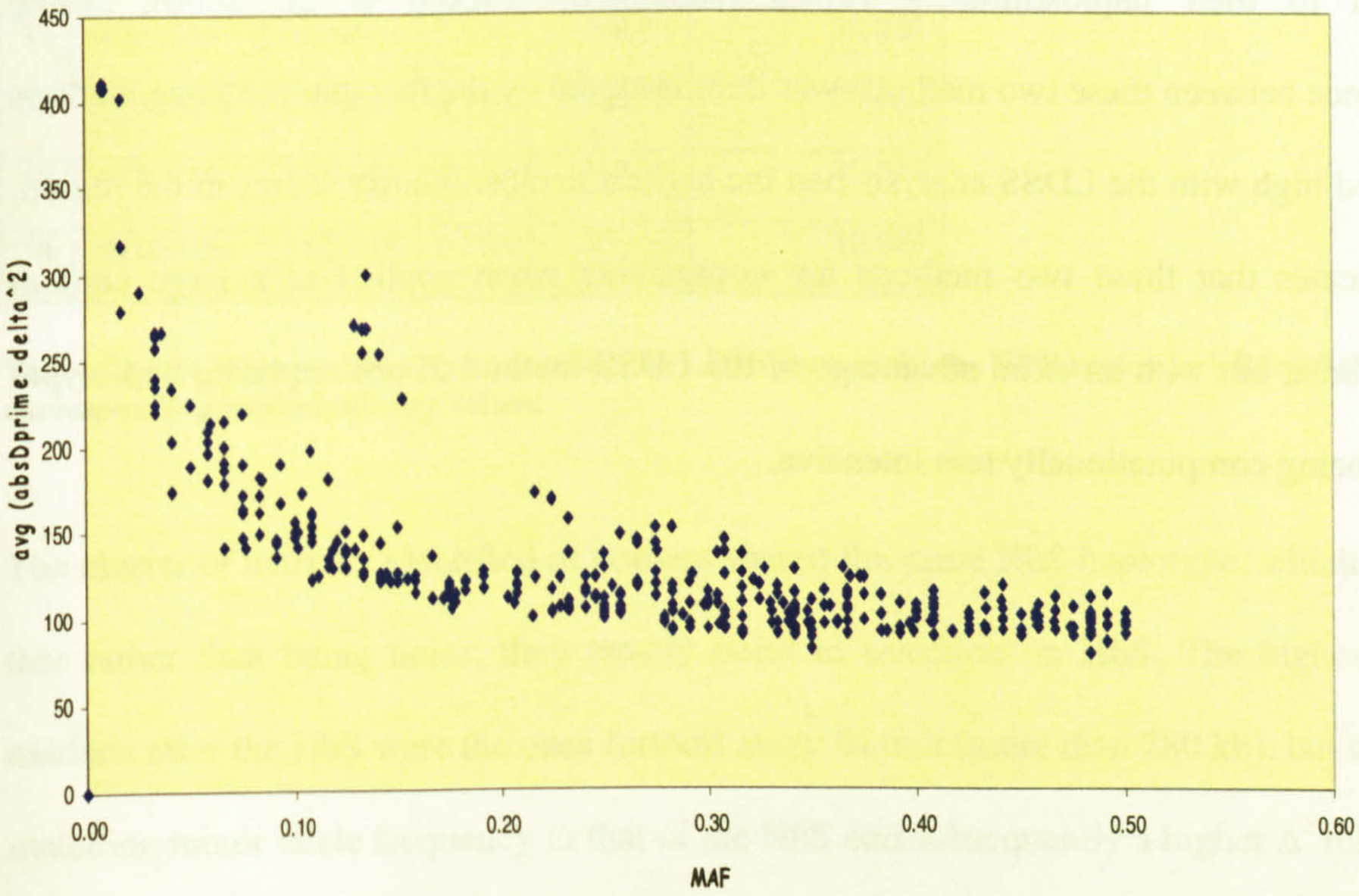


Figure 6.5.5.1: A scatter plot of the correlation between minor allele frequency MAF (on the x axis) and LD-summary statistic (y axis) for each marker in the 1.1Mb region anchored on the HbS in YRI.

For the resulting scatter plot a smoother line was fitted using a generalized additive model, this model divides the data into overlapping intervals and for each interval tries to fit the best regression line that describes that section of the data. 95% confidence intervals are then calculated for the regression lines. For each data point the residual value is calculated by taking the vertical difference between the data-point value on the Y axis and its fitted value on the smoother. The absolute values of the residuals are then standardized by dividing by the standard deviation to determine the statistical significance of the deviation of each data point from the general distribution of the rest of the data.

A few markers stood out as obvious outliers not conforming to the correlation curve distribution. These are shown in table 6.5.5.1. When the LDSS values of these markers were compared to their haplosimilarity values (Hanchard, Rockett et al. 2006), strong concordance between these two methods was demonstrated by the fact that the same markers that scored high with the LDSS analysis had the highest haplosimilarity scores in the region. This indicates that these two methods are comparable when applied to a large enough genomic area, but with an extra advantage of the LDSS method of not requiring haplotypic data and being computationally less intensive.

id	*rs#*	*position*	*MAF*	*LDSS*	*haplosimilarity*
1	rs7951605	4883681	0.225	167.844	14.422
2	rs7945056	4887272	0.225	167.844	14.902
3	rs334(HbS)	5204808	0.135	298.845	32.942
4	rs7119428	5266389	0.216	171.788	15.903
5	rs4519119	5267214	0.216	171.788	15.185
6	rs2213170	5267342	0.216	171.788	15.34
7	rs2226952	5271463	0.225	167.136	15.295
8	rs417425	5474593	0.133	253.674	26.777
9	rs392296	5475237	0.153	226.419	23.529
10	rs414154	5484881	0.129	269.418	31.202
11	rs393044	5500846	0.136	266.72	30.234
12	rs1391614	5505802	0.133	266.72	30.594
13	rs317775	5510959	0.142	252.78	30.015
14	rs7934354	5530215	0.233	156.158	16.926

Table 6.5.5.1: Markers identified as outliers by LDSS analysis in the HBB region in YRI and their corresponding haplosimilarity values.

The cluster of markers identified as outliers shared the same HbS haplotype, which indicates that rather than being noise, they mostly point to selection on HbS. The highest scoring markers after the HbS were the ones furthest away from it (more than 280 kb), but they had a matching minor allele frequency to that of the HbS and subsequently a higher Δ^2 relationship with it. For the other markers in the cluster, they were of a higher minor allele frequency than that of the HbS marker. They had a high lD'l but low Δ^2 values with the HbS marker and they also shared but to a lesser extent the same HbS haplotype.

To further confirm that the cluster of unusual LDSS values observed resulted from selection on the HbS allele, I analysed areas of the same size, both upstream and downstream of the 1.1Mb area anchored on HbS. I found that the clustering of signals noted in the core 1.1Mb region, was absent from the two flanking regions (Figure 6.5.5.2 and 6.5.5.3). This indicates that all the signals around the β-globin region point towards or result from selection pressure on the HbS allele and the clustering effect of signals is due to the presence of a single high frequency extended haplotype carrying the HbS allele.

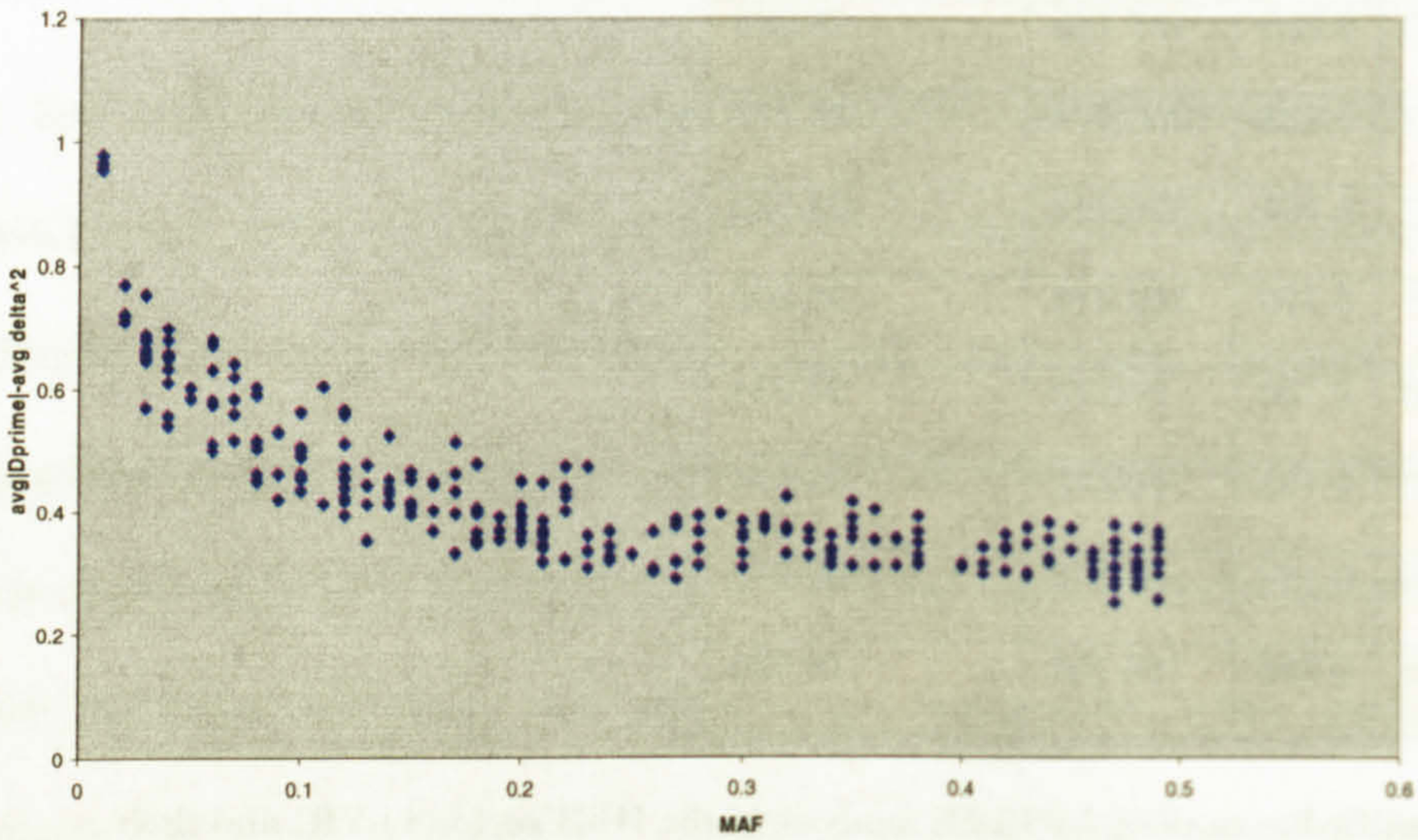


Figure 6.5.5.2: LDSS analysis of a 1Mb region upstream of the 1.1Mb anchored on the HbS marker.

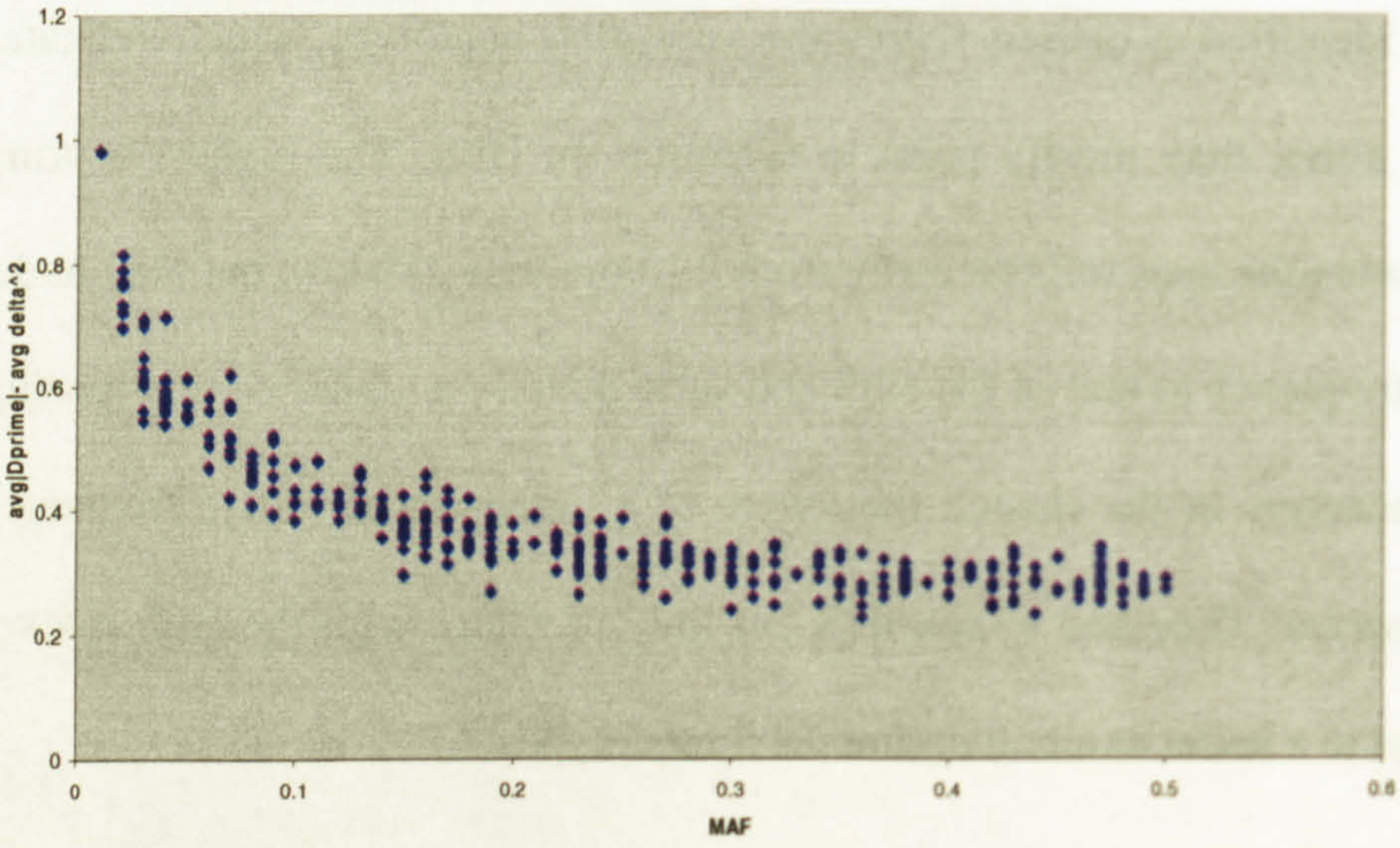


Figure 6.5.5.3: LDSS analysis of a 1Mb region downstream of the 1.1Mb anchored on the HbS marker.

I also carried out the LDSS analysis on the whole of chromosome 11 with a sliding window of size 380 markers (approximately 1Mb) and shift of 180 markers. This analysis showed windows around the HbS mutation to have a cluster of outlier markers. This clustering effect appears to be only around the HbS variant on chr11, a result which supports the previous high frequency haplotype analysis and could be used as an alternatively simpler method to scan the genome for regions under positive selection.

Another piece of evidence suggesting that positive selection is responsible for the observed LDSS signals was facilitated by the knowledge of which hapolotypes carry the HbS mutation. I removed the seven identical haplotypes containing the HbS allele out of data of the 1.1 Mb region around HbS, and then reanalysed the data with the LDSS metric. The results showed disappearance of the cluster of signals observed previously in the region when the HbS haplotypes were taken out (Figures 6.5.5.4).

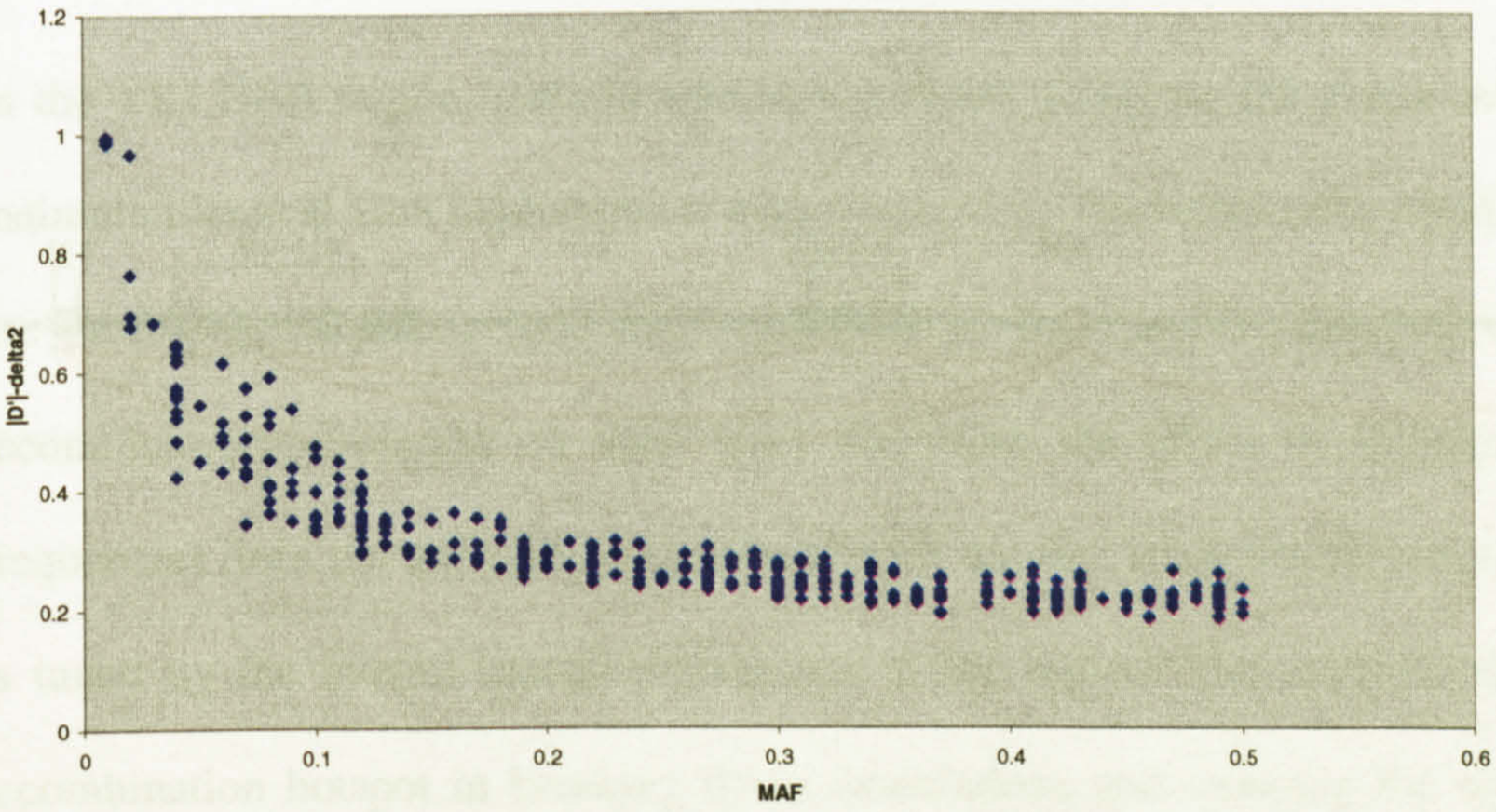


Figure 6.5.5.4: LD-summary statistic carried out in the 1.1Mb region centred on the HbS after removing the 7 identical HbS haplotypes. LDSS scores are displayed in the y axis and Minor Allele Frequencies on the x axis. Figure shows disappearance of outliers when compared with figure 6.5.5.1.

The positions, MAFs, LD and haplotypic relationships between markers displaying signals of positive selection in a genomic region could potentially be used to better predict location and MAF of the functional SNP. In the 1.1Mb region anchored on the HbS SNP, the markers that had the highest LD-summary statistic and haplosimilarity scores were those closest to the causal variant (HbS) in terms of allele frequency and haplotypic relationship, but not necessary in terms of physical distance. Therefore, it might be reasonable to generalize that in a given genomic region with a cluster of markers displaying signals of being positively selected, the causal variant (if not one of the markers typed) is likely to be of a similar MAF and shares the same haplotypes with those of the highest scoring markers.

On the other hand, based on that information alone, no assumptions could be made about the position of the functional variant in the region. For that purpose the LD relationship between all members of the cluster could give clues about the position as can be seen from the example of the genomic area around the HbS in the YRI (Figure 6.5.5.5).

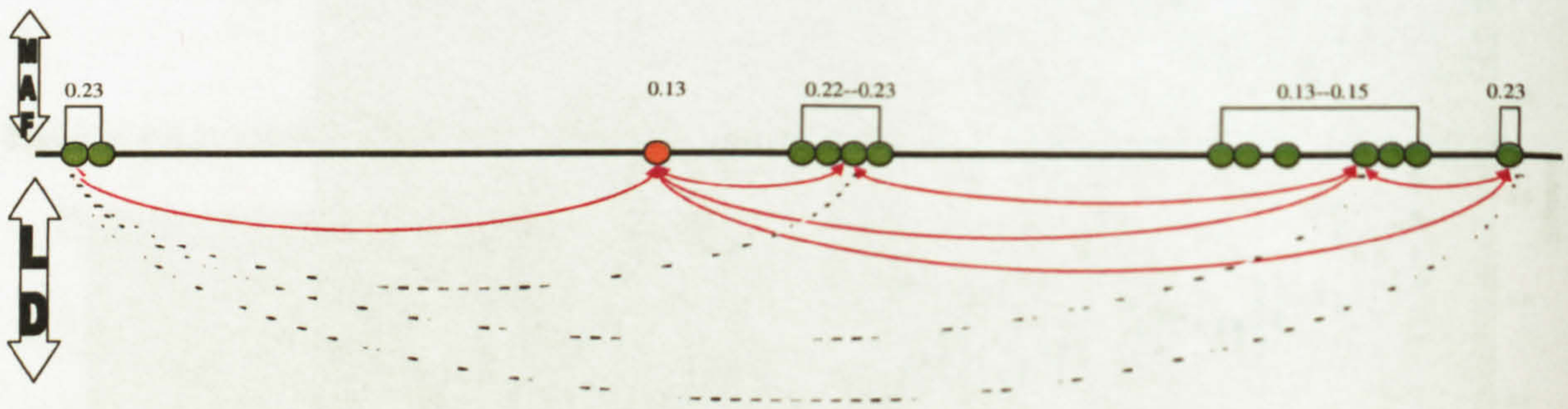


Figure 6.5.5.5: A schematic drawing of MAF and LD relationships between outlier markers in the LD-summary statistic analysis in a 1.1Mb region around the HbS in the YRI. HbS is shown as an orange circle and other markers as green circles on the horizontal axis. MAF are indicated on the top half of the figure above each group of markers. The LD relationships represented on the bottom half of figure with red lines indicating strong LD and interrupted line indicating weak LD.

In the schematic figure above, it is shown that all cluster members had strong LD relationships (represented by red lines) with the HbS SNP and with each other within the group each side of the HbS, but not among the two groups of markers on either side of it. If the causal variant is not known and a guess had to be made about which one it is, it is more likely to be the one with strong LD with all the other members of the cluster. In the case where the functional marker is not typed then the closest estimation of its position would be in the area where the LD breaks down between members of the cluster.

6.6. Discussion

6.6.1. Quantifying the temporal relationship between recombination and the effects of selective sweeps on haplotype frequency distribution

In the YRI HBB region, malaria selection pressure acting on the sickle-cell variant helped maintain identical HbS haplotypes at high frequency. This effect was equal on both sides of the HbS polymorphism (600 kb) regardless of the number, position and intensity of recombinational hotspots on each side. Therefore the effect of selection on haplotype frequencies does not seem to be correlated with the fine scale recombination rate but rather is tuned by the overall recombination rate in the region. The reported effect of a single recombination hotspot in breaking down associations and reducing the signal of selection around HbC (Wood, Stover et al. 2005) does not hold true for this analysis. This might be due to the weaker selective pressure of the HbC mutation compared with the HbS.

6.6.2. Strengths and limitations of the method

Using the extended-high-frequency-haplotype method is a simple and quick way to highlight a particular genomic region as a candidate of natural positive selection, as well as defining the boundaries of that region for further analysis. The simplest form of such downstream analysis could be the LD-summary statistics for markers within that region's boundaries. This type analysis will probably not be as informative without defining the unit size by the initial scan.

This analysis helped identify and describe the genomic scale over which selective sweeps could have an effect. With the extended-high-frequency-haplotype method, detecting positive selection in genomic regions could be achieved without regard to whether the causal variant was typed or not. Consequently, there is less emphasis on the choice of markers, density and spacing unlike other methods (like LRH and haplosimilarity) which rely on the ability of a marker to tag the causal SNP by being in high LD with it and thus making marker choice and density of essential importance. Using data for all SNPs in a genomic region in the order of a megabase makes this method robust to marker choice and density variation when compared to the above mentioned methods. The consistency in finding the high frequency extended haplotype in the face of variable marker density, and chance element in choice and ascertainment of typed markers, gives this method an advantage by decreasing the rate of false negatives when looking for signals of positive selection.

This property may make this method useful for genome wide case control studies on a large number of individuals with a modest marker coverage that will not necessary tag all the untyped markers, a thing which is logistically difficult to achieve either due to limitations in resources, technology or an over-fitting problem in marker choice which in most instances rely on an imperfectly transferable SNP-tagging sets between populations and studies.

Using the correlation between haplotype frequencies and recombination rate as a test to look for selective sweeps will miss those regions with no or very little recombination rates if they were acted on by positive selection. As it stands now this analysis is a conservative way to scan for selection, in the sense that it would only have power to pick up areas with incomplete selective sweeps which are relatively recent and did not yet reach fixation, due to the underestimation of recombination rates in regions with complete sweeps.

The choice of window size, that would be optimal for identifying genomic regions under selection, warrants some consideration for different datasets. At very small window sizes, most haplotypes will be of a high frequency which makes them indistinct from the selected haplotype. At the other extreme of very big windows, all haplotypes would be distinct from each other leading to a failure to pick the selected haplotype. Between these two extremes of distribution uniformities, all the possible signals with different effect sizes could potentially be identified by running the analysis with different window sizes.

In the YRI I fixed the window size to 360 markers, which roughly corresponded to 1Mb. I chose this size because it is optimised to selective sweeps of a similar or greater magnitude to that observed for the HbS in YRI.

6.6.3. Gene ontology

The percentage of genes involved in mediating inflammatory and immune responses in areas picked up by the scan was five times more than that of the proportion of immune genes in the whole genome (10% vs 2%). The higher preponderance of immune genes in these regions is highly suggestive of them being selectively important.

6.6.4. Why there are more sweeps in CEU than YRI

The migration of modern humans out of Africa into new environments was accompanied by genetic adaptations to emergent selective forces. An interesting feature of my data is that the majority of deviations are not shared between the two population samples, suggesting that local adaptation has played an important role in recent human evolutionary history. Consistent with this observation, several possible examples of local adaptation in humans have previously been reported (Rana, Hewett-Emmett et al. 1999; Hollox, Poulter et al. 2001; Tishkoff, Varkonyi et al. 2001; Currat, Trabuchet et al. 2002; Fullerton, Bartoszewicz et al. 2002; Gilad, Rosenberg et al. 2002; Hamblin, Thompson et al. 2002; Rockman, Hahn et al. 2003).

The highest signal in the CEU sample was found in the region of chromosome 2 that contains the LCT gene which previously was found in Northern European populations to have very high frequencies of the lactase persistence allele (LCT*P) (Hollox, Poulter et al. 2001), which allows digestion of fresh milk throughout adulthood. It is widely accepted that strong selection has driven LCT*P to high frequency in Northern Europeans, beginning sometime after the domestication of animals approximately 9,000 years ago (Hollox, Poulter et al. 2001; Bersaglieri, Sabeti et al. 2004).

The stronger signatures of selection in the European-derived population may reflect the exposure of non-African populations to novel and evolutionarily recent selective pressures (e.g., unique dietary, climatic, and cultural environments) as modern humans migrated out of Africa and spread throughout the world. In contrast, the Yoruba population may have experienced fewer evolutionarily recent selective forces. Theoretical studies have demonstrated that the power to detect a selective sweep is generally greatest if it occurred less than approximately 0.1 Ne generations ago (i.e., approximately 20,000-25,000 years ago

(Kim and Stephan 2000; Przeworski 2002), which is consistent with the hypothesis that signatures of selection in European-Americans reflect recent selective events. Glinka et al. (Glinka, Ometto et al. 2003) found that European-derived populations of *Drosophila melanogaster* demonstrated abundant evidence for recent selective sweeps, whereas African populations did not, which is strikingly similar to my results.

6.6.5. The genome-wide approach

One way to overcome the confounding effects of population history is by empirically comparing the pattern of variation at a candidate locus with the genome-wide pattern estimated from a set of neutral markers that have been typed in the same individuals or populations. In contrast to demographic processes, which affect the entire genome, natural selection affects specific functionally important sites in the genome.

The availability of large-scale genomic data has created new challenges and opportunities, especially in allowing for more outlier analyses. However, the availability of genomic data is not the final answer the fundamental problem that population-level demographic processes and selection are confounded. Many demographic processes, such as certain types of population subdivision, may increase the variance in the statistics used to detect selection. Certain demographic models are, therefore, more likely than other models to produce outliers. The outlier approach in population genetics does not solve the problem that a postulated signature of selection, inferred from population genetic data, may instead be the product of complicated demographics. Therefore more downstream analysis to try and locate the functional variant is required.

6.6.6. Towards locating the functional variant

The fact that the high-frequency-extended-haplotype method identifies genomic regions in the order of megabases as possible targets of natural selection has its pros and cons. It lends robustness to identifying positive selection signals without the need to specifically type the functional SNP as long as the haplotype diversity is correctly captured, overcoming any biased estimation introduced by SNP ascertainment. The possibility of capturing the true haplotype diversity becomes more likely the bigger the area studied at any one time. Conrad et al (Conrad, Jakobsson et al. 2006) found that the longer the haplotypes considered, the truer the estimate of their heterzygosity. They also found that the same underlying haplotypes are likely to be observed, regardless of which SNPs are studied, over a long enough genomic region.

The boundaries of the region over which subsequent analysis could be carried out, can be demarcated very effectively by identifying the extent of a long haplotype of unusually high frequency. Without this area demarcation some analysis like the LD- summary statistic analysis will not give any coherent or useful result.

The challenge of this method stems from its source of strength. It becomes more of a challenge to pinpoint the functional variant, the larger the genomic area over which the search has to be conducted. A few markers in the regions identified by the extended high frequency haplotype method are expected to have unusually extensive LD values because of their close correlation with the high frequency selected haplotype. To try and identify these markers I defined a metric based on LD and called it the LD-summary statistic (LDSS). It is a quick method which proved to give comparable results to those of the best methods at the moment, when used within the proper pre-defined genomic area context.

The Expectation Maximization (EM) algorithm was used to calculate LD values between markers in regions spanned by an extended high frequency haplotype. The EM algorithm was chosen because it is much less computationally intensive than haplotype-based LD calculations and it gives comparable results. The gain in computational time and effort over using the haplotype based methods, overbalance the slight loss of accuracy in LD estimates. Certainly for the purposes of this analysis the EM algorithm is quite adequate as long as the bias in the estimation is consistent, since it is the overall average LD summary of each marker rather than the detailed pair-wise LD values that are being used.

There were eight haplotypes with the minor alleles of all of the outlier markers in the LDSS analysis and all the eight haplotypes were carrying the HbS allele. This close haplotypic relationship between causal variant and other members of cluster of markers with selection signals in the region could be utilized to limit the depth of the search for the causal variant by marking a few haplotypes out for further genotyping effort and analysis.

All markers that had high signals on both the haplosimilarity and LDSS analysis were found to be highly correlated with the haplotypes carrying the HbS allele. This suggests that rather than being noise, the signals they show mostly point to selection on HbS.

To confirm this, the haplotypes containing the HbS were taken out of the analysis with a resultant marked reduction of signals at those markers that stood out before. No reduction in their signals was noted when a similar number of randomly chosen haplotypes was taken out of the analysis. At the same time for the other markers which demonstrated no significant signals before taking the HbS haplotypes out, there was hardly any change to their haplosimilarity signals. This observation supports the claim that most of the haplosimilarity signals in the HBB region are the result of the selection on one marker and they could all point towards a single selective pressure on the HbS allele. So rather than considering these

signals the result of stochastic events, an indication of the insensitivity of the metric used or a problem to be circumvented, they could be utilized to characterise genomic regions undergoing selection and to pin point the position of the functional polymorphism.

In order to achieve the above goal (the utilization of selection signals at specific markers to identify genomic regions under selection), I had to prove first that there is a tendency of selection signals to be clustered in genomic regions undergoing selective sweeps. The clustering effect, observed very clearly in the 1.1Mb region anchored on the HbS, was found to be absent from the same sized flanking regions using the LD-summary statistic analysis.

The whole chromosomal LDSS analysis carried out across chromosome 11 with a sliding window of approximately 1Mb in size, showed the clustering effect to be only around the HbS mutation in chr11. This result supports the previous high frequency haplotype analysis and could be used as an alternatively simpler method to scan the genome for regions under positive selection.

Positive selection signals at several individual markers could all collectively be pointing towards selection on a single untyped functional marker. In the HBB region unless the functional marker (HbS) is typed it could be easily overlooked in the search for selection signal in the genome using LRH and derived metrics. The very essential act of defining the extent of the area where to look for a cluster of signals, regardless of the metric used, makes this approach more robust than the approach of trying to find a single truly selected marker.

6.6.7. Method correlation with EHH and haplosimilarity

In chromosome 11, the region that showed unusually high frequency long haplotype with the extended-high-frequency-haplotype analysis was the same region that demonstrated clustering of signals using other haplotype based analysis like the LRH and haplosimilarity.

The same cluster of markers that stood out in the LRH and HS analysis was picked up by the LDSS analysis, which shows that this method lines up very closely to the other haplotypic based analyses if the overview of the clustering effect of a selective sweep is considered rather than the individual markers scores which are at best hard to interpret against the background noise.

The most important insight from my analyses is highlighting the scale over which signals of selection are most effectively detected, and giving other methods of looking for natural selection context by considering all the members of a cluster of signals in a genomic region to be telling the same story.

6.6.8. Applying the method to genome wide association studies with less marker density and more individuals than the HapMap data

To explore whether using high density markers in a small area yields the same results as using widely spaced markers in a bigger region, an LD-summary analysis was carried out in a 150kb region around the HbS marker in YRI, using the HapMap release 20 data which is about five times denser than the data used for my analysis so far. 200 markers were analysed and it was found that the HbS marker is hardly distinguishable from the background noise because its average LD value is not that much different from other markers distribution (Figure 6.6.8). This leads to the conclusion that to first define the boundaries of the area over which the LDSS can be carried out is of utmost importance in successfully identifying the positive signal from the background noise.

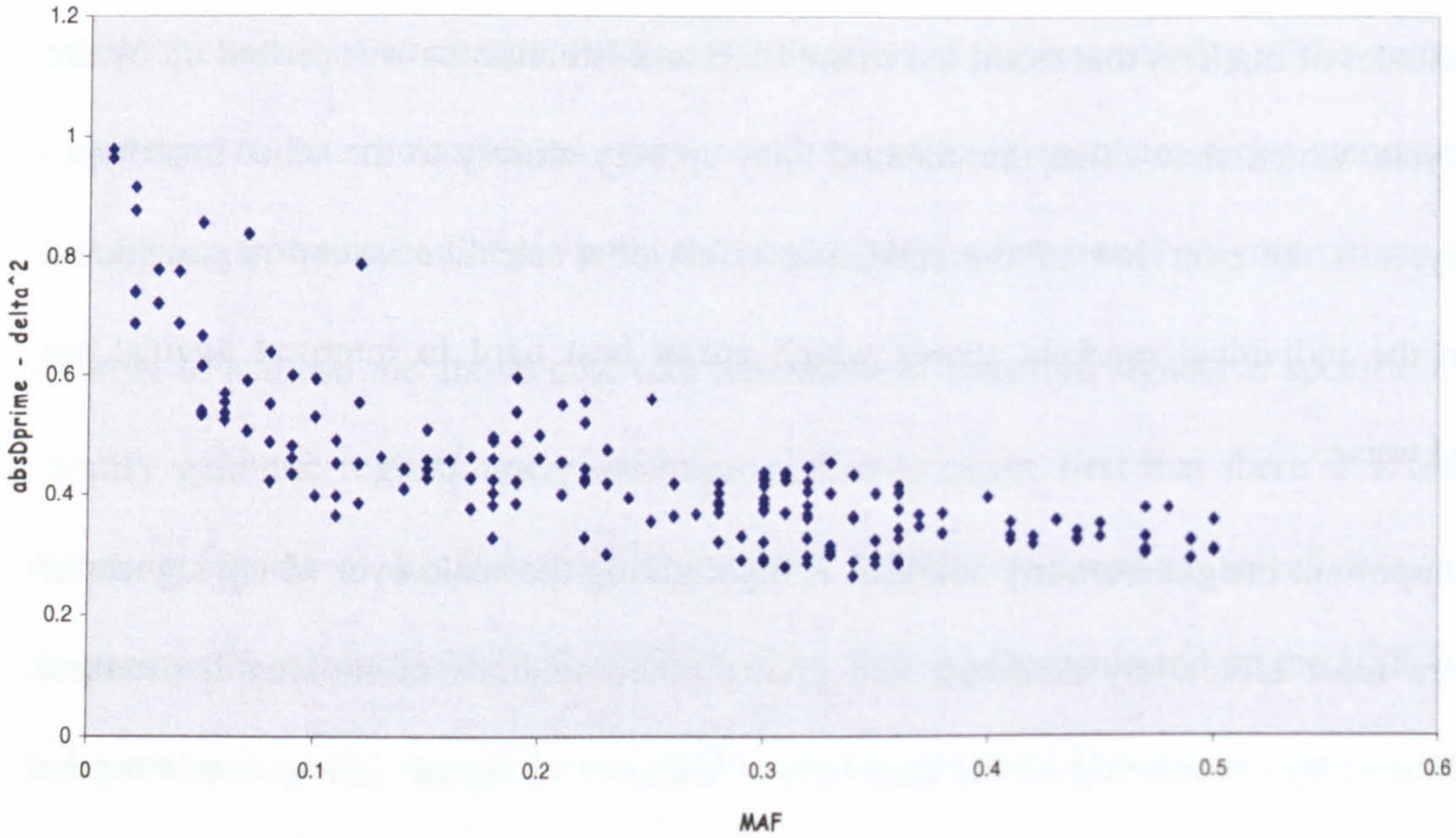


Figure 6.6.8: LD-summary statistic analysis of a 150 kb region anchored on the HbS in YRI. 200 markers from HapMap release 20 were used.

The density of the markers in the 1.1 Mb region around HbS was reduced to half that in phase 1 of the HapMap, by taking every other marker's genotypes out and then re-phasing the genotypic data. The HbS haplotypes were still distinguished from others by their high frequency over that distance.

In a recent study (Conrad, Jakobsson et al. 2006) it was shown that the bigger the window considered; the more powered the genotyped SNPs to capture the haplotypic heterozygosity in the area as measured by microsatellites.

My method is suited to genome wide association studies with much more individuals typed and less marker density than that used in phase 1 of the HapMap project, because it will probably capture the same haplotypic diversity with less marker density. Furthermore, it is likely that the difference in frequency between selected and other neutral haplotypes in the same region will become more prominent with the larger number of individuals typed.

To summarize, the LD-summary analysis highlighted the clustering effects of selective sweeps, and the potential for using the LD and haplotypic relationships between these cluster markers to better predict location and minor allele frequency of the functional SNP. Within the boundaries of the genomic region under selection and the full scale of the sweep, the LD-based analysis is as good as other computationally intensive haplotype-based methods at picking up selection signals. As well as coming up with comparable results, it gave context to the interpretation of signals obtained by other methods.

6.7. Conclusion

The same striking pattern I previously observed in the Sudanese samples was mirrored in the Yoruba HBB region with a high frequency HbS haplotype that extended to 1.2 Mb, and which clearly stood out when compared with the rest of chromosome 11. Genome-wide scan identified 23 genomic regions in the YRI, and 32 regions in the CEU, with unusually extended high frequency haplotypes. These regions were enriched for immune genes, suggesting they might have been targeted by positive selection for their important biological functions. The highest signal in the YRI genome was that from the HBB region. In the CEU genome the most remarkable signal mapped to the 2q21.3 region within which the LCT gene resides. Both of these regions have extensive literature suggesting their selective advantage.

The genome-wide approach I employed in the HapMap data to look for extended-high-frequency- haplotypes could be useful for highlighting genomic areas that are good candidates of positive selection. Since this approach is relatively robust to the choice and density of markers and could potentially pick up signals of selection even if the selected variant is not genotyped, it might prove useful for genome-wide case control studies.

Therefore in the next chapter I will apply the same approach developed here to a larger dataset generated by the MalariaGEN project. This dataset consists of genome-wide genotyping data of severe malaria cases from the Gambia and their parents as controls.

Chapter 7:

Genome-wide detection of malaria-related natural selection by applying an extended-high-frequency haplotype method to case-control data from the Gambia.

7.1. Abstract

The extended-high-frequency-haplotype method introduced in the previous chapter showed promising results, suggesting its utility in highlighting genomic regions that might be candidates of positive selection. Applying it to a real-life genome-wide case-control data presents the opportunity to further develop and validate the method, and to identify genomic regions where positive selection might have played a role. Additionally, this analysis could draw disease-specific inferences that might aid the search for malaria resistance/susceptibility genetic variants.

Here I analyse the results of applying this method to a case-control study of Gambian children with severe malaria and their parents as part of the Malaria Genomic Epidemiological Network (MalariaGEN) project. 2632 chromosomes were analysed for 585,350 SNPs (total number of genotypes 770,320,600). Overlapping windows of 1cM size and 0.1cM shift were run across the 22 autosomes. Data from all chromosomes was then combined to carry out a genome-wide statistical assessment where the upper 2.3% of the data points were highlighted and further explored for their genic content. This analysis was carried out separately for the malaria cases and controls.

7.2. Objectives

- Scan the genome of Gambian malaria cases and controls by the high-frequency-extended-haplotype method to recognize putative genomic regions under positive selection.
- As a proof of principal, investigate whether there is a selection signal at the HbS locus, and if so, if the signal is more prominent in the controls indicating malaria-specific selective pressure.
- Investigate whether there is a selection signal in the MHC region, and if so, if the signal is malaria related.
- Look for other prominent signals in the genome.

7.3. Introduction

Human genetic diversity has been shaped, in part, by infectious diseases, and one of the more prominent examples of this is the impact of selective pressure exerted by malaria. In areas where malaria is endemic, those individuals who carry resistance-conferring genes stand a better chance of surviving childhood and contributing to the population's genetic pool. This has long been recognized when the theory of natural selection was substantiated for thalassemias, sickle cell anaemia, and other red blood cell disorders (Flint, Harding et al. 1998). Haldane's insight in1949, that the geographical distribution of haemoglobinopathies reflected malarial selection, provided numerous candidate genes for studies over the next four decades.

Twin studies and heritability estimates have subsequently confirmed the influence of host genetics, which was shown to be most pronounced in children (Jepson, Banya et al. 1995; Rihet, Abel et al. 1998; Mackinnon, Mwangi et al. 2005).

The importance of genes regulating immune responses to malaria was demonstrated by the finding of HLA associations with resistance to severe malaria (Hill, Allsopp et al. 1991). Polymorphism in the promoter of another MHC gene, tumour necrosis factor TNF, was found to affect the risk of cerebral malaria (McGuire, Hill et al. 1994). However it has been surprisingly difficult to detect an influence of HLA and other major histocompatibility complex genes on the magnitude of immune responses to malarial antigens in field studies. In general, cellular immune responses to malaria antigens show marked heterogeneity in specificity, type and magnitude; the relative importance of MHC polymorphism and other genetic factors in accounting for this heterogeneity has been unclear (Hill, Jepson et al. 1997).

A number of clear associations of genetic polymorphisms with the altered risk of malaria disease have been presented and supported by a significant body of scientific evidence in the past, mainly from candidate gene approaches (Fortin, Stevenson et al. 2002). Most of these association studies have targeted immune loci, thus highlighting the importance of immune processes in malaria. Some examples are loci encoding MBL, CD36, CD40 ligand, IFN-γ, IL4 and the p40 subunit of IL12 (Kwiatkowski 2000).

But despite the large number of field studies over many years, knowledge on the key targets and mechanisms of protective immunity is still remarkably limited, with many studies giving negative or contradictory results. Numerous studies reporting apparent associations between polymorphisms and outcome of malaria infection have been published in the last 10 years,

but there are concerns that many of these associations may be spurious because of the small size of the studies and low levels of population matching.

Recently, data availability from huge international initiatives like the human genome sequence, the international HapMap project, and large-scale multicenter studies of exposed populations using immunogenetic studies of polymorphisms (such as those of the MalariaGEN research network) (MalariaGEN 2008), along with the advent of new methods for high throughput, high resolution genotyping are leading to an explosion in genetic epidemiology, which will pave the way for better understanding of mechanisms of malaria protective immunity and the rational development and evaluation of future vaccines and novel therapies.

Theory and analytical approaches used to detect signatures of natural selection in the human genome

Natural selection, which can be defined as the differential contribution of genetic variants to future generations, is the driving force of Darwinian evolution. Identifying regions of the human genome that have been targets of natural selection is an important step in clarifying human evolutionary history and understanding how genetic variation results in phenotypic diversity, it may also facilitate the search for complex disease genes. Rather than detecting selection by observing its ongoing dynamics, population genetic approaches aim to establish whether or not observed extant patterns of genetic variation would be unlikely in the absence of selection.

The traditional way of identifying targets of adaptive evolution has been to study a few loci that one hypothesizes a priori to have been under selection. This approach is complicated because of the confounding effects that population demographic history and selection have on patterns of DNA sequence variation. Technological advances in high-throughput DNA

sequencing and single nucleotide polymorphism genotyping have enabled several genome-wide scans of natural selection to be undertaken.Using a genome wide distribution of any statistic can potentially tease out the signature of selection from a background signature of demographic history. Because while specific demographic events, such as population expansions, bottlenecks, and subdivision of populations will potentially affect variation genome wide. On the other hand, natural selection is expected to have locus-specific effects.

The deluge of large-scale catalogues of genetic variation has stimulated many genome-wide scans for positive selection in several species. Recently the Phase II HapMap data have been used to identify genomic regions that show evidence for the influence of adaptive evolution, primarily through extended haplotype structure indicative of recent positive selection. Using two established approaches, namely the LRH test and the Integrated Haplotype Score (iHS) test (Sabeti, Reich et al. 2002; Voight, Kudaravalli et al. 2006), approximately 200 regions with evidence of recent positive selection from the Phase II HapMap were identified. These regions include many established cases of selection, such as the genes HBB and LCT, the HLA region (Sabeti, Varilly et al. 2007).

7.4. Materials and Methods

7.4.1. Samples

I was privileged to have been given the opportunity to work with data generated by the Malaria Genomic Epidemiological Network (MalariaGEN) project, in order to test the applicability of my method and assist the MalariaGen analysis group in exploring new ways

of analyzing the data. By applying the extended-high-frequency-haplotype method to the data, the opportunity was presented for further development and validation of the method.

MalariaGEN is an international research Network formed by researchers from more than 20 countries in Africa, Asia, North America and Europe, in order to better understand mechanisms of protective immunity against malaria. The project involves collaboration between investigators in 25 institutions which are all not-for-profit research institutes or universities (MalariaGEN 2008).

Here I describe the results when the extended-high-frequency-haplotype method was applied to the Gambian dataset of MalariaGen Consortial project 1 (CP1). The following is a description of how this data was collected, generated and processed by the MalariaGEN project.

Consortial Project 1 (CP1): Whole genome association study of severe malaria:

The aim of this investigation was to identify sequences of DNA (single nucleotide polymorphisms (SNPs) which are associated with susceptibility or resistance to malaria. MalariaGEN partners participating in this project contributed samples of DNA from well-defined cases of severe malaria with ethnically-matched controls. Phase 1 of the project involved screening hundreds of thousands of SNPs for association with susceptibility or resistance to severe malaria in Gambian children with severe malaria and their parents. Unlike case-control analysis, family based association analysis is not confounded by population stratification which could increase the rate of false positive and false negative results.

Clinical Samples:

In this study cerebral malaria was defined as a Blantyre coma score of < 3, persisting for > 30 minutes after cessation of a transient seizure or after correction of hypoglycaemia, in a child with asexual forms of *P. falciparum* on blood film and no other evident cause of coma. Severe malaria anemia was defined as packed cell volume < 15%, or hemoglobin < 5 (or 6) g/dl with asexual forms of *P. falciparum* on blood film.

Whole genome amplification:

DNA samples were whole genome amplified using f29 multiple displacement amplification (MDA) with REPLI-g^{TM} 625S reagents based on instructions from the manufacturer (MSI Inc, New Haven). The quality and quantity of DNA was assayed in each sample prior to amplification with PicoGreen.

The DNA samples were in TE (10mM Tris-HCl pH 7.5, 1mM EDTA) and the concentration was 20ng/mL in a total volume of 10mL. Amplified DNA samples were re-assayed using PicoGreen, normalized to 250 ng/mL, and loci representation QC was performed by Taqman assay on two loci. All f29MDA DNAs selected for genome-wide genotyping resulted from reactions with a minimum of 5 ng input genomic DNA. The quality of the amplified DNA was further assessed by assaying 30 SNPs across the genome using Sequenom iPlex. Call rates and imputation on these DNA samples were assessed in the following study (Teo, Inouye et al. 2008).

7.4.2. Genotyping

Extracted DNA specimens were sent to Panos Deloukas lab in the Sanger Institute for genotyping services using Illumina 650Y chip (Ilmn650K) SNP array. This chip is based on the HumanHap550 panel which is a whole-genome genotyping panel of 555,352 SNPs that

was constructed to effectively tag CEU (European) and CHB + JPT (Asian) sample populations. A majority of the SNPs were selected by tagging the more than 2 million common HapMap SNPs, but the panel also includes variation types that have been found to be overrepresented in diseases such as nonsynonymous SNPs, SNPs in the MHC region, SNPs in commonly reported CNV regions, and mitochondrial SNPs. Because individuals with African ancestry have distinct and lower levels of LD compared to those with European or Asian ancestry, another 100,000 common YRI (African) tag SNPs were added to increase coverage of the YRI samples for the HumanHap650Y panel. Eberle et al estimated the genome coverage and power for the HumanHap650Y panel (Eberle, Ng et al. 2007).

Genotype calling:

The kind of data volumes that were generated by high throughput genotyping platforms, created a need for an efficient and accurate automated genotype calling software.

A genotype calling algorithm for the Illumina BeadArray genotyping platforms was developed and applied to the data by Teo et al (Teo, Inouye et al. 2007) who formalized a calling strategy with a number of features that are specifically designed for the BeadArray genotyping technology. They introduced a model-based approach to call genotypes for the Illumina BeadArray platforms and this has been implemented within an Expectation-Maximization framework in the program *Illuminus*.

Illuminus made more concordant calls and resulted in a smaller number of SNPs which are excluded on the basis of per-SNP call rates than other available genotype calling algorithms. This improvement was found to be significantly more substantial for DNA samples which have undergone whole-genome amplification.

Data Quality Control (QC):

SNPs QC: Those SNPs which had more than 5% of their genotypes missing were excluded. Also SNPs with excess HWD, and Mendelian errors were excluded.

Samples QC: Excluded were the samples with more than 5% missingness, excess heterozygosity, any duplicates. For the trios all Mendelian discrepancies were set to missing before phasing.

658 trios and 585,350 SNPs passed the quality control, and this was the dataset on which I conducted my analysis.

7.4.3. Haplotypic phasing

Phasing was carried out by Kerrin Small using the **Phamily-PHASE** and **PHASE** ***version2.1*** software packages (Stephens, Smith et al. 2001; Stephens and Donnelly 2003).

The genotyping data was partitioned into 50 SNP segments conditional on the presence of at least one unambiguous SNP in each segment. If this condition was not fulfilled, more SNPs would be added until this condition was met, up to a maximum of 200 SNPs.

The ***Phamily*** algorithm was first used on the trios to determine the phasing that can unambiguously be inferred from the pedigree data. Afterwards, **PHASE** was used to statistically infer the rest. In this process any missing data was imputed.

Finally, the different segments were patched together to create the haplotypes along the whole length of the chromosome, for all autosomes.

Files with the phased haplotypes for the Gambian trios were provided, one file for each chromosome, in the following format:

Father id
Haplotype transmitted to the child
Haplotype not transmitted to the child
Mother id
Haplotype transmitted to the child
Haplotype not transmitted to the child
Child id (case)
Haplotype transmitted from father
Haplotype transmitted from mother

Along with each haplotype file there was a legend file with data on the typed markers, their chromosomal location based on NCBI Genomic build 35, and their genetic distance from the start of the chromosome. Estimates of genetic distances were those calculated from and provided by phase 2 HapMap.

7.4.4. Selection analysis

To start with, analysis was carried out separately for each of the different chromosomes. Perl and R scripts running on a UNIX platform were developed and used to carry out the analysis (Appendix 3). The same procedure was conducted for each chromosome as follows:

- Firstly, two files were generated from data contained in the file with the phased haplotypes of the Gambian family trios. The haplotypes of the parents were divided into those that were transmitted to the children, representing the severe malaria cases; and the rest of the paternal haplotypes that were not transmitted to the children, representing the controls. The resulting two groups of haplotypes were saved in two separate files.

- Before starting analysing the haplotypes, windows were defined beforehand based on the data from the accompanying legend input file, where all the typed markers were listed with their physical and genetic distances away from the chromosomal start. The coordinates of the markers defining the boundaries of windows were saved in addition to information about the number of markers in each window, and the windows' sizes in terms of genetic and physical distances.

- Overlapping windows covering the whole length of a chromosome were fixed to a maximum of a 1cM in size. When there were no typed markers at the exact window limit, the last marker position in a 1cM range from the window start, was taken as the window's end position. Therefore, in spite of pre-fixing the genetic distance of windows there was still some variability in this statistic.

- To create a sliding window effect, the start positions of successive windows were shifted by incrementing the previous window starting position by 0.1cM (when no markers were typed at the required exact position, the closest marker before that position was taken as the window starting point). Thus, the overlap created between consecutive windows was roughly 0.9cM.

- After all the markers in all windows across the chromosome were defined, including first and last markers' physical and genetic distance coordinates, the haplotypic information for these markers were retrieved from the haplotype file. The search for unusually high frequency extended haplotypes was carried out once for the cases and then for the controls as described below.

- In each window, the number of copies (frequency) of each distinct haplotype in that window was counted, and then those haplotypes and their copy number were ordered by frequency from the highest to the lowest. The number of copies of the most frequent haplotype in the window was recorded.
- The same procedure was carried out for all the overlapping windows across each chromosome.

Data from all chromosomes were combined to carry out a genome-wide statistical assessment and analysis as follows:

- Firstly, all windows with less than 0.8 cM in size or less than 50 typed markers were filtered out of the analysis.
- The absolute frequencies of the haplotypes with the maximum number of copies were plotted for all windows across all chromosomes.
- A generalized additive model (R function *gam*) was used to correlate the number of typed markers in a window with its highest haplotype frequency. Thus controlling for the variability in the number of typed markers that affects the resolution of haplotype ascertainment.
- This statistical model partitions the data into overlapping segments, then fits a statistical mean that best describes the data in each respective segment. The means are then smoothed to create an overall mean for the whole of the data.
- For every data point its residual value (that is its vertical distance away from the mean) is calculated.

- The standard deviation is calculated. Then every point's residual value is divided by the standard deviation to give the standardized residual value for that data point.
- All data points with standardized residual values above two or more are taken to be significant, which means that the upper 2.3% of the data points were considered to be enriched for selection signals and were then further explored for their genic content in the malaria cases and controls.

7.4.5. Gene set analysis

WebGestalt which is a "**WEB**-based **GE**ne **S**e**T** **A**na**L**ysis **T**oolkit" was used to carry out analysis on gene sets produced from the above analysis. WebGestalt incorporates information from different public resources and provides an easy way for biologists to make sense out of large sets of genes. It enables biologists to manipulate integrated information and find patterns that are not detectable otherwise. **WebGestalt** is designed for functional genomic, proteomic and large scale genetic studies from which high-throughput data are continuously produced. It currently works for human and mouse (http://bioinfo.vanderbilt.edu/webgestalt/).

7.5. Results

In total there were 2632 chromosomes analysed for 585,350 SNPs, which makes for a total number of 770,320,600 genotypes. This volume of data required the development of a number of programming scripts to automate the handling and analysis of this dataset. Overlapping windows of a fixed size (1cM), and shift (0.1cM) were used to analyse the

haplotype data in all 22 autosomes of the Gambian trio data. As part of the analysis, information about window statistics (window size in bp, window size in cM, window start and end coordinates, and number of typed markers) were gathered and recorded in the initial stage. The average window size in terms of physical distance was 761613 bp. The average window size in terms of genetic distance was 0.93 cM. The average number of markers typed per window was 166 markers. The variation in the genetic distance was not eliminated by pre-fixing the window sizes to 1 cM due to markers spacing variability. Figure 7.5.1 shows the frequency distribution of window sizes in cM for all windows across all chromosomes.

Before analysing the haplotype frequencies, any window which was 0.8 cM or less in size was eliminated from the analysis. This made for 5% of data being filtered out. The logic behind this is that the smaller the genetic distance considered, the less recombination events considered, which means the higher likelihood of haplotypes with high frequencies. A factor which would bias analysis standardised for genetic distance by increasing the likelihood of false positives.

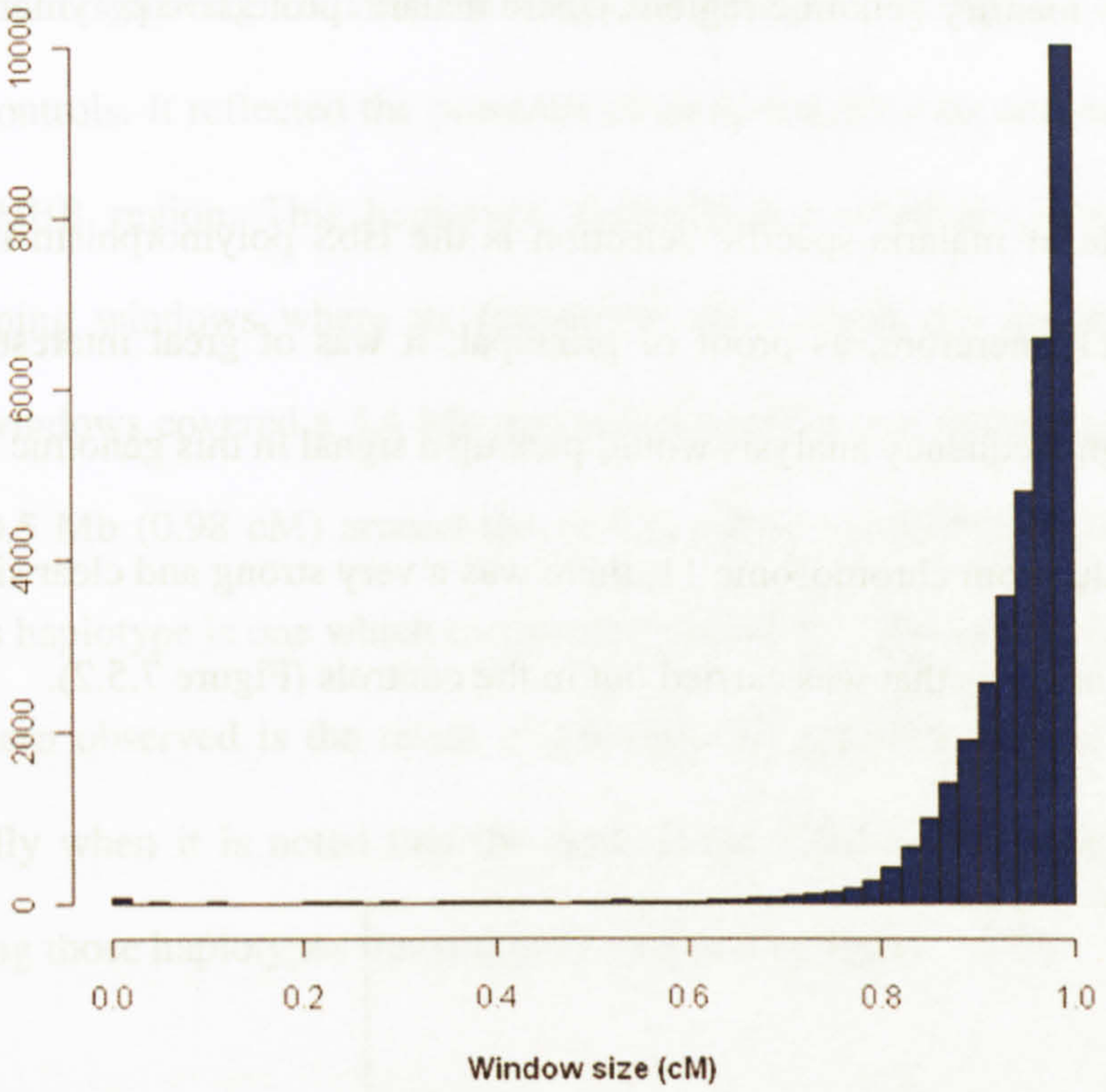


Figure 7.5.1: Frequency distribution of window size statistic. Data is for all windows across all autosomes in the Gambian trios. Shown on the x axis the bins of window size in cM. On the y axis the frequency of each bin.

Another potential bias could arise from the variability of the number of markers typed in each window. Although the analysis is standardized to account for the number of markers per window, initially I chose to exclude all the windows with 50 or less typed markers. 4% of windows across the genome had 50 or less typed markers, but about half of those were also 0.8 cM or less in size. Therefore, the combined proportion of windows filtered out of the analysis was 7%.

Each chromosome was analysed twice; one time with the haplotypes that were transmitted from parents to children (representing the severe malaria cases), and a second time with the haplotypes that were not transmitted from the parents (representing the controls). Therefore, the analysis carried out in the cases can potentially highlight genomic regions where genetic variants might be involved in malaria susceptibility, and the analysis carried out in the

controls could potentially identify genomic regions where malaria protective polymorphisms reside.

The bench mark example of malaria-specific selection is the HbS polymorphism in HBB region in chromosome 11, therefore, as proof of principal, it was of great interest to see whether the extended-high-frequency analysis would pick up a signal in this genomic region.

When looking at the results from chromosome 11, there was a very strong and clear signal in the HBB region with the analysis that was carried out in the controls (Figure 7.5.2).

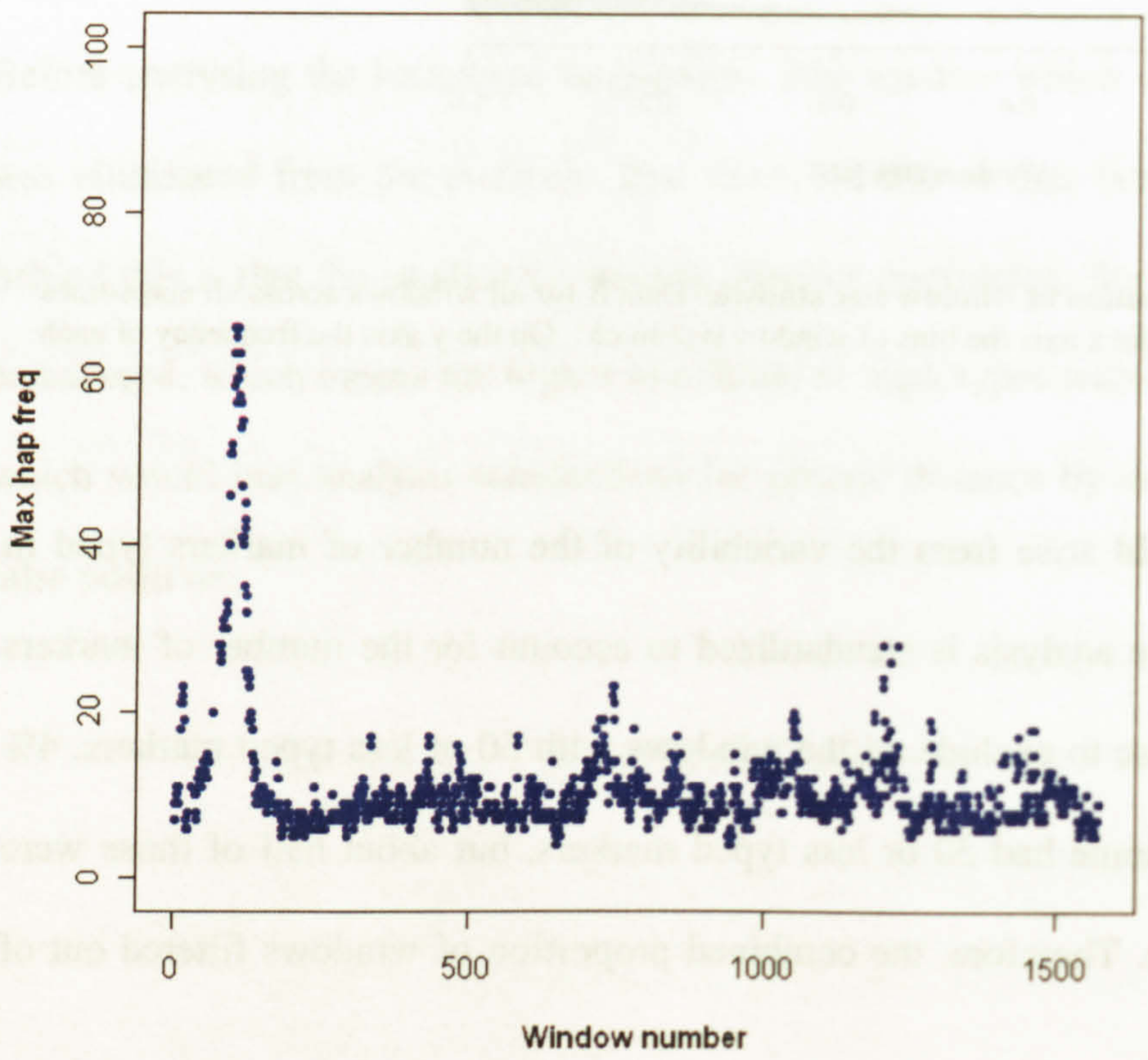


Figure 7.5.2: Analysis of untransmitted haplotypes (malaria controls) of chromosome 11. Scatter plot of maximum haplotype frequencies in windows of 1 cM size and 0.1 cM shift along chr11. Only windows more than 0.8 cM in size and with more than 50 typed markers are shown in figure. On the x axis, windows are arranged according to their position along chr11, with the p arm on the left hand side. On the y axis, the number of copies of the most frequent haplotype in each window are shown.

The signal originating from the HBB region constituted the highest signal in chromosome 11 in the controls. It reflected the presence of an unusually long and high frequency haplotype in the HBB region. This haplotype maintained a relatively high frequency over many overlapping windows where its frequency was outside the genome-wide 95th percentile. These windows covered a 3.4 Mb genomic area. The maximum frequency of 66 copies was in the 0.5 Mb (0.98 cM) around the HbS position (11:4795740_11:5309382). I confirmed that this haplotype is one which carries the HbS allele. This makes for a very strong case that the pattern observed is the result of selection on the HbS allele in the Gambian controls, especially when it is noted that the peak at the HBB region disappears completely when analysing those haplotypes transmitted to the cases (Figure 7.5.3).

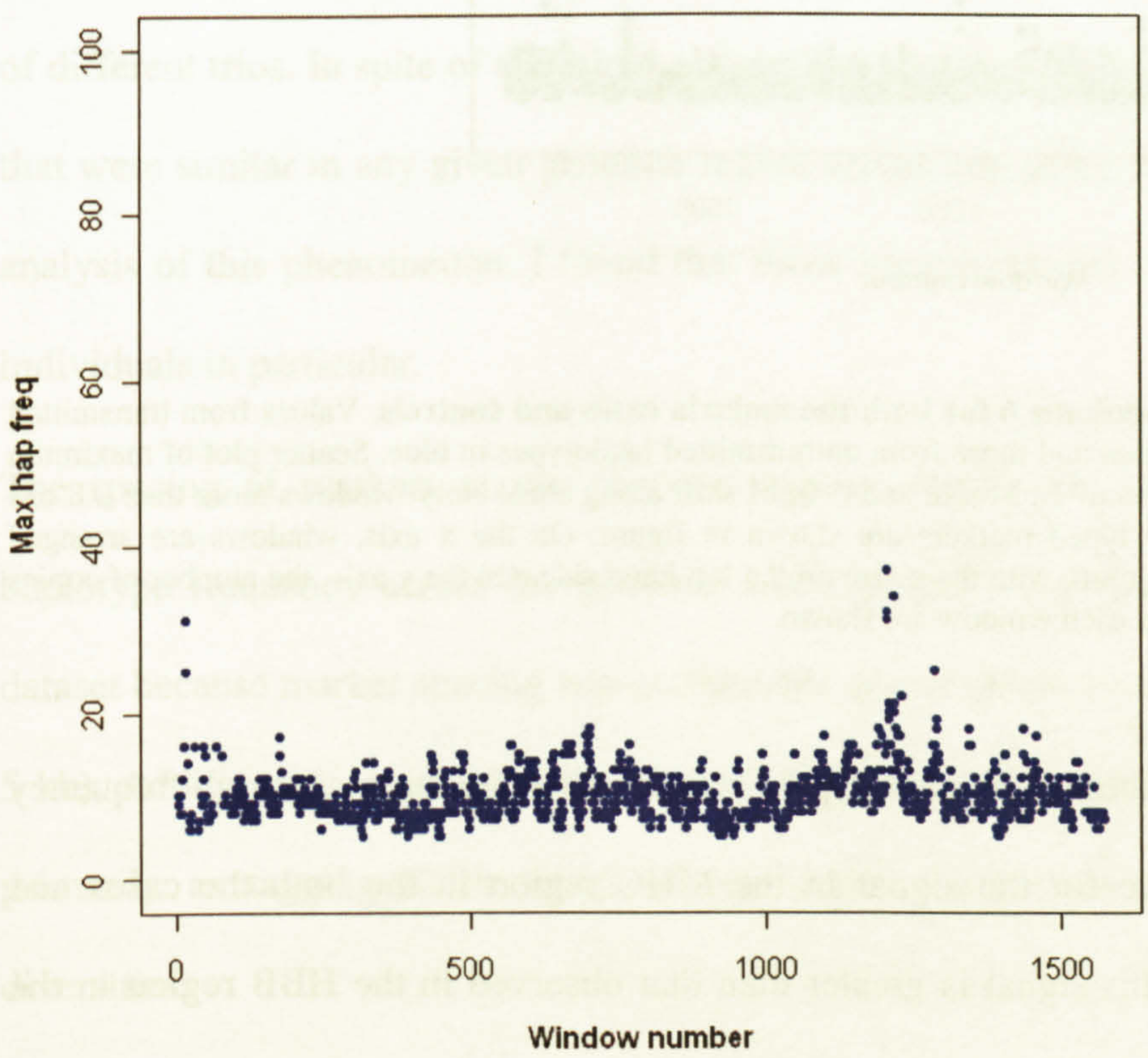


Figure 7.5.3: Analysis of transmitted haplotypes (malaria cases) of chromosome 11. Scatter plot of maximum haplotype frequencies in windows of 1 cM size and 0.1 cM shift along chr11. Only windows more than 0.8 cM in size and with more than 50 typed markers are shown in figure. On the x axis, windows are arranged according to their position along chr11, with the p arm on the left hand side. On the y axis, the number of copies of the most frequent haplotype in each window are shown.

There were cases when the peaks were observed both in the transmitted haplotypes (cases) and untransmitted haplotypes (controls). The most striking example of that is the genomic area in chromosome 6 where the MHC and other immune genes are located (Figure 7.5.4).

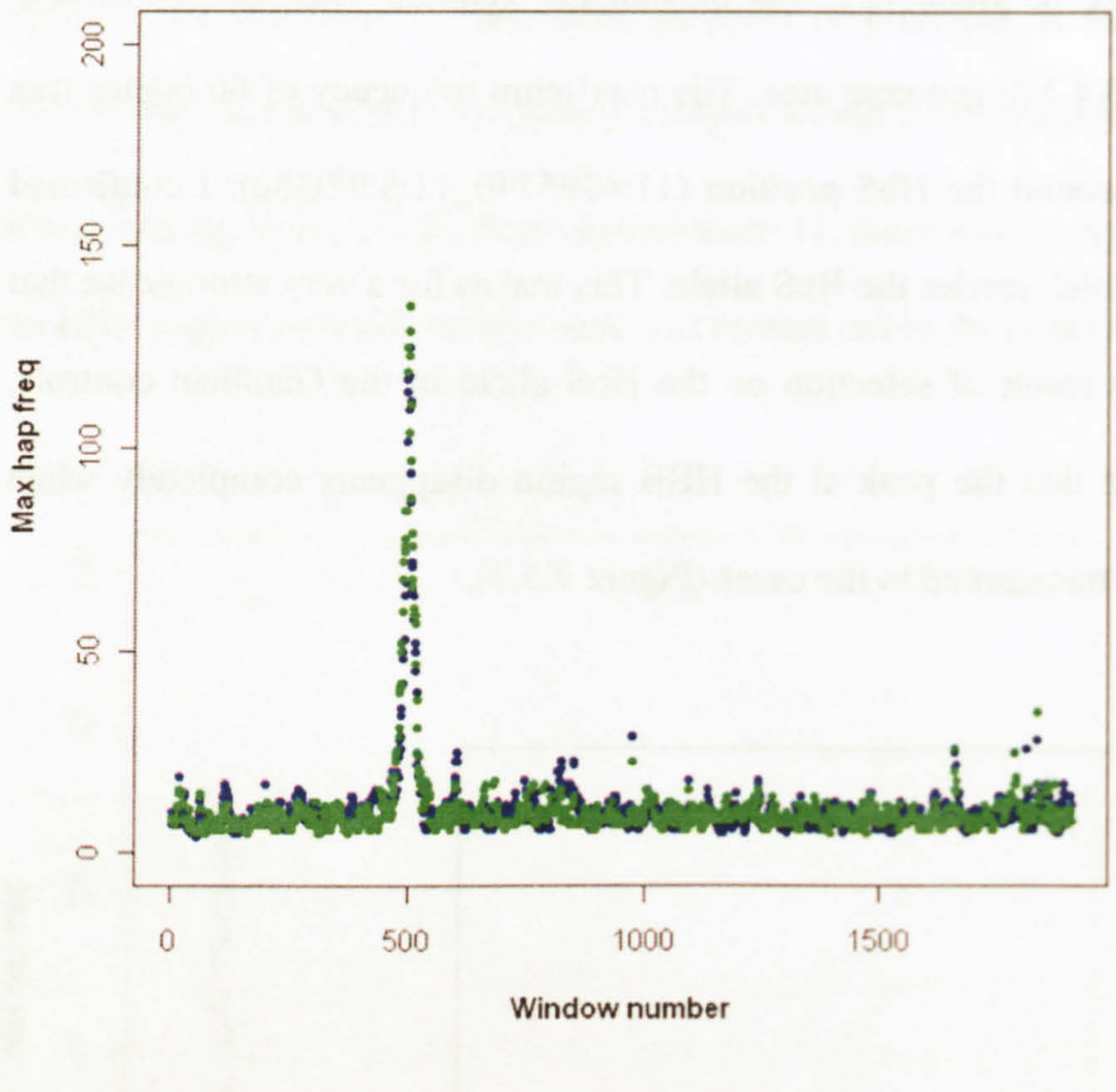


Figure 7.5.4: Analysis of chromosome 6 for both the malaria cases and controls. Values from transmitted haplotypes are represented in green and those from untransmitted haplotypes in blue. Scatter plot of maximum haplotype frequencies in windows of 1 cM size and 0.1 cM shift along chr6. Only windows more than 0.8 cM in size and with more than 50 typed markers are shown in figure. On the x axis, windows are arranged according to their position along chr6, with the p arm on the left hand side. On the y axis, the number of copies of the most frequent haplotype in each window are shown.

From looking at the sequences of the haplotypes, I found the same unusually high frequency haplotype to be responsible for the signal in the MHC region in the both the cases and controls. The strength of this signal is greater than that observed in the HBB region in the malaria controls.

Due to the many immunologically important genes in this region, it is not implausible that the signal observed might be due to the effects of a very strong natural positive selection force acting on the region, but not necessary due to malaria.

Data from all windows across all chromosome were compiled together to make a comprehensive genome-wide analysis and determine significance level across the whole genome. The pooling of the data from different chromosomes was a statistically feasible thing to do since the factor that most determine haplotype frequencies in a region, namely the recombination rate, was accounted for in the analysis. Another factor that might possibly influence the variability in haplotypes frequencies across the genome is sample relatedness. It was excluded since the Gambian sample consisted of unrelated trios, and there was no evidence from the genetic data to suggest there were undetected kinships between members of different trios. In spite of the initial observation that there was a baseline of 10 haplotypes that were similar in any given genomic region across any given chromosomes, upon further analysis of this phenomenon, I found that those haplotypes did not belong to any group of individuals in particular.

The spacing of markers across genomic regions might also influence the variability in haplotype frequency across the genome. But I judged it not to be a major factor in this dataset because marker spacing was comparable across chromosomes (one marker every 4 to 5 kb). Additionally, markers in the Illumina 650Y chip (Ilmn650K) SNP array were chosen primarily as tagging SNPs which means their spacing was correlated with haplotype diversities.

The number of markers typed in each window however can greatly influence the observed haplotype frequency in that window. The smaller the number of markers per window, the less resolution of distinct haplotypes and consequently the higher likelihood of observing a

haplotype with high frequency in that window. In figure 7.5.5 it is clear that there is greater variability in the maximum haplotype frequencies across windows with smaller number of typed markers.

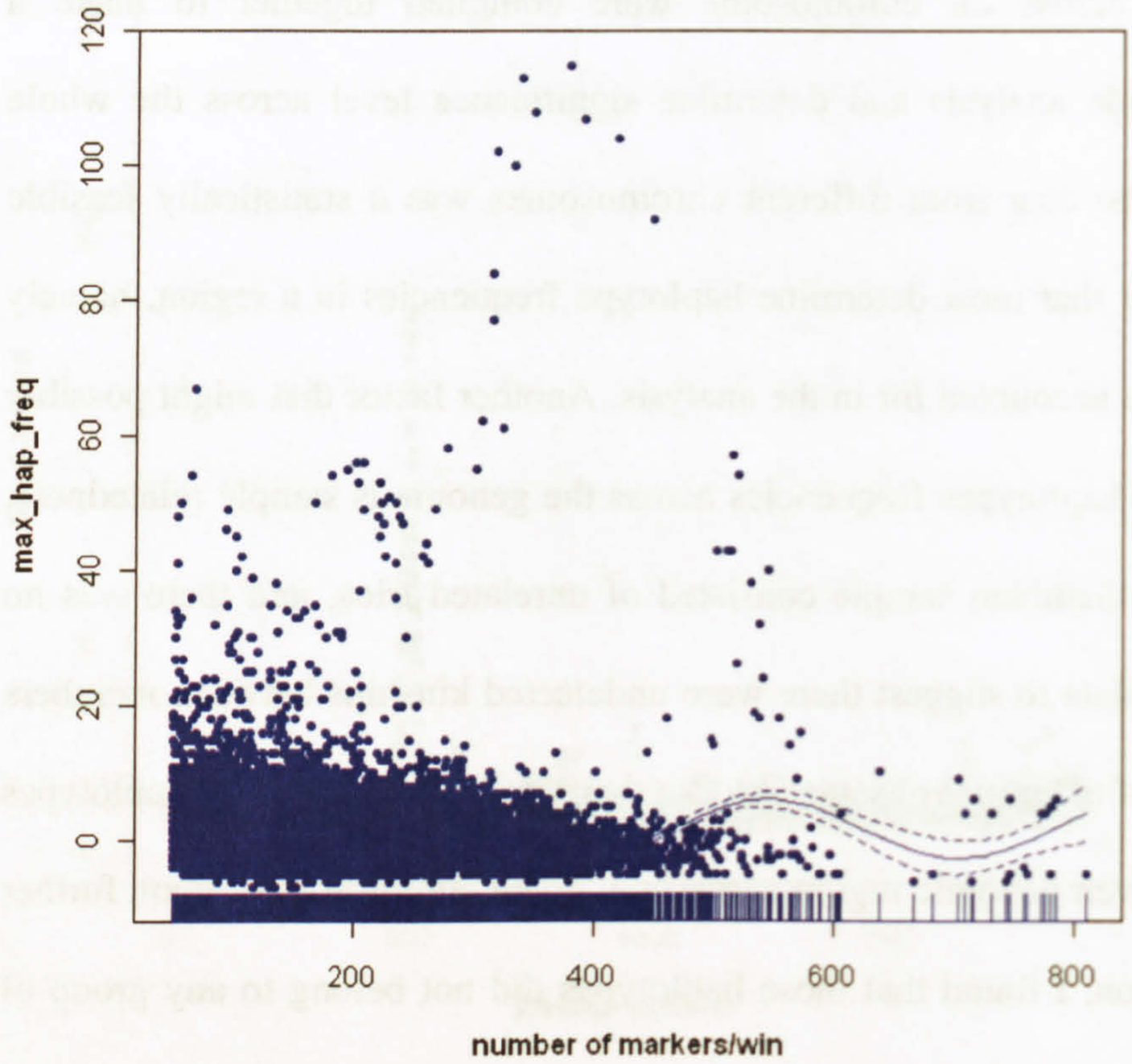


Figure 7.5.5: Correlation between the maximum number of haplotype copies in a window and the number of typed markers. Windows are of 1cM size and 0.1cM shift. Each data point on the scatter plot represent a window with its number of typed marker on the x axis, and the copy number of its most frequent haplotype on the y axis.

The way I chose to tackle this issue was first to exclude all windows that had 50 or less typed markers. For the remaining data the variability in window statistics with the number of typed markers was factored in to the analysis by statistically correlating the number of markers typed per window and its highest haplotype frequency. A generalized additive model (R function *gam*) was used which partitions the data to overlapping segments then fits a statistical mean that best describes the data in each respective segment. The means were then smoothed to create an overall mean for the whole of the data.

For every data point its residual value (its vertical distance away from the mean) is calculated. The standard deviation is calculated. Then every point's residual value is divided by the standard deviation to give the standardized residual value for that data point. All data points with standardized residual values above two or more were taken to be significant. This resulted in the upper 2.3% of the data considered to be outliers and warrant further exploration for their potential biological significance.

In figure 7.5.6 the standardised residual values for filtered data are displayed by chromosomal order for the severe malaria cases.

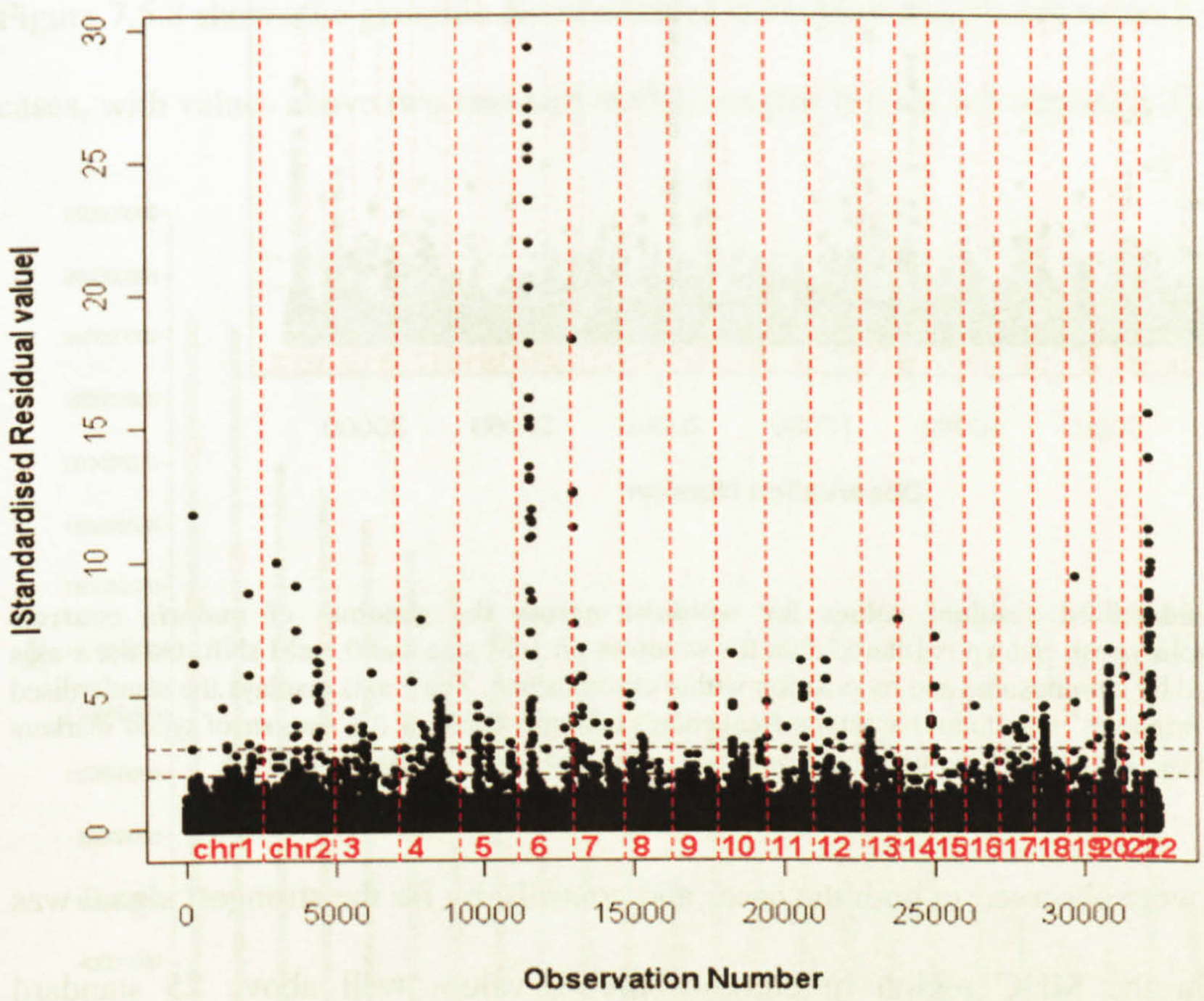


Figure 7.5.6: Standardised residual values for windows across the genomes of malaria cases (transmitted haplotypes). Shown is filtered data for windows on 1cM size and 0.1 cM shift. On the x axis windows are ordered by chromosome and by position within chromosome. The y axis displays the standardised residual values for windows' maximum haplotype frequencies taking account of the number of typed markers in the window. The two grey horizontal lines represent the second and third standard deviations.

This analysis was carried out using the haplotypes transmitted from parents to children. Each chromosome was done independently then the data was pooled for all chromosomes and statistical significance determined. The same was done for the controls on the parental haplotypes not transmitted to children with severe malaria (Figure 7.5.7).

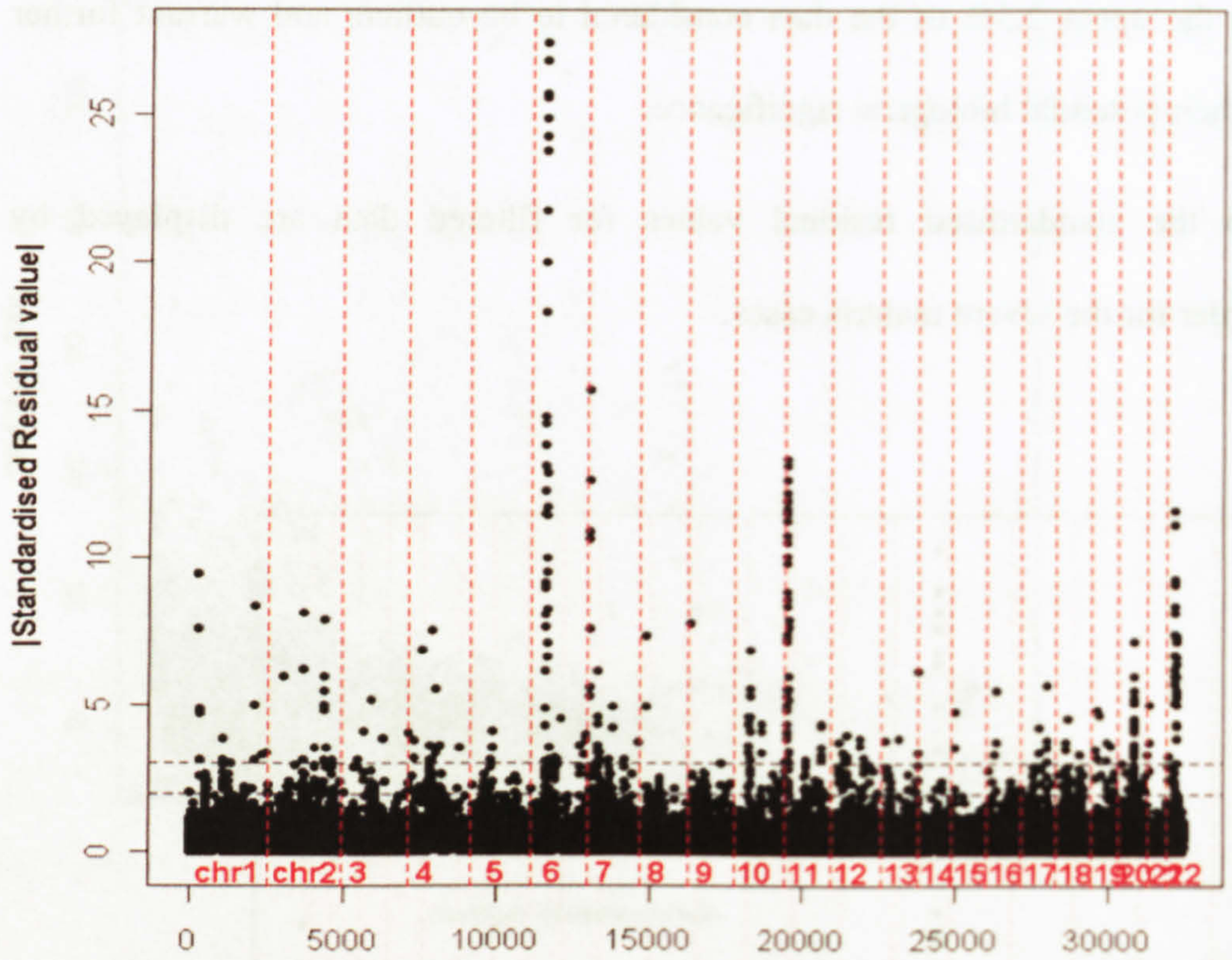


Figure 7.5.7: Standardised residual values for windows across the genomes of malaria controls (untransmitted haplotypes). Shown is filtered data for windows on 1cM size and 0.1 cM shift. On the x axis windows are ordered by chromosome and by position within chromosome. The y axis displays the standardised residual values for windows' maximum haplotype frequencies taking account of the number of typed markers in the window. The two grey horizontal lines represent the second and third standard deviations.

For signals that were observed in both the cases and controls, by far the strongest signal was that observed in the MHC region in chromosome 6, values well above 25 standard deviations of the genome wide average. In the controls, the highest signal exclusive to the controls originated in the HBB region of chromosome 11 (11:3524620-11:6871891), which was the second strongest genome wide after the MHC signal, with values above the 13th standard deviation. A number of additional regions with evidence of unusually long high

frequency haplotypes were identified in the cases and controls. Although their statistical significance might be weaker than regions mentioned above, still they might be plausible candidates for positive selection. Experiences in other infectious diseases have shown that weaker signals may as well lead to the identification of relevant genetic variants (Ogura, Bonen et al. 2001).

Certainly in their entirety, outlier regions in this analysis are likely to be enriched for biologically important genes under selection. Further exploration of these and other regions identified by similar analysis and localization attempts for causal variants is being carried out by the analysis group of the MalariGen project.

Figure 7.5.8 shows the genomic distribution of the regions, exclusive to the severe malaria cases, with values above two standard deviations (for full list see appendix 4).

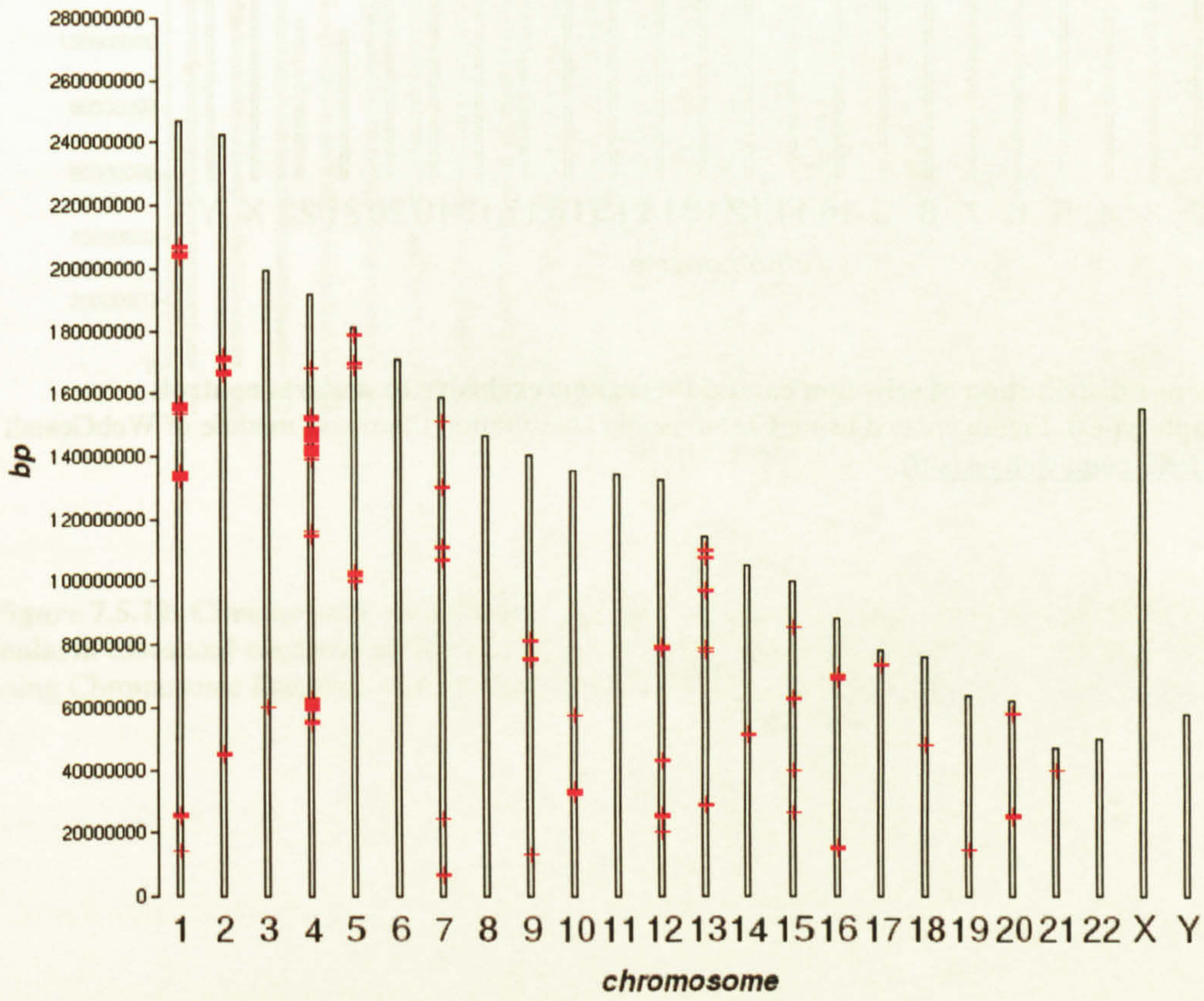


Figure 7.5.8: Genomic distribution of selection candidate regions exclusive to malaria cases (transmitted haplotypes). Figure created using Chromosome Distribution Chart sub-module of WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/).

Figure 7.5.9 shows the genomic distribution of the regions, exclusive to the controls, with values above two standard deviations (for full list see appendix 4).

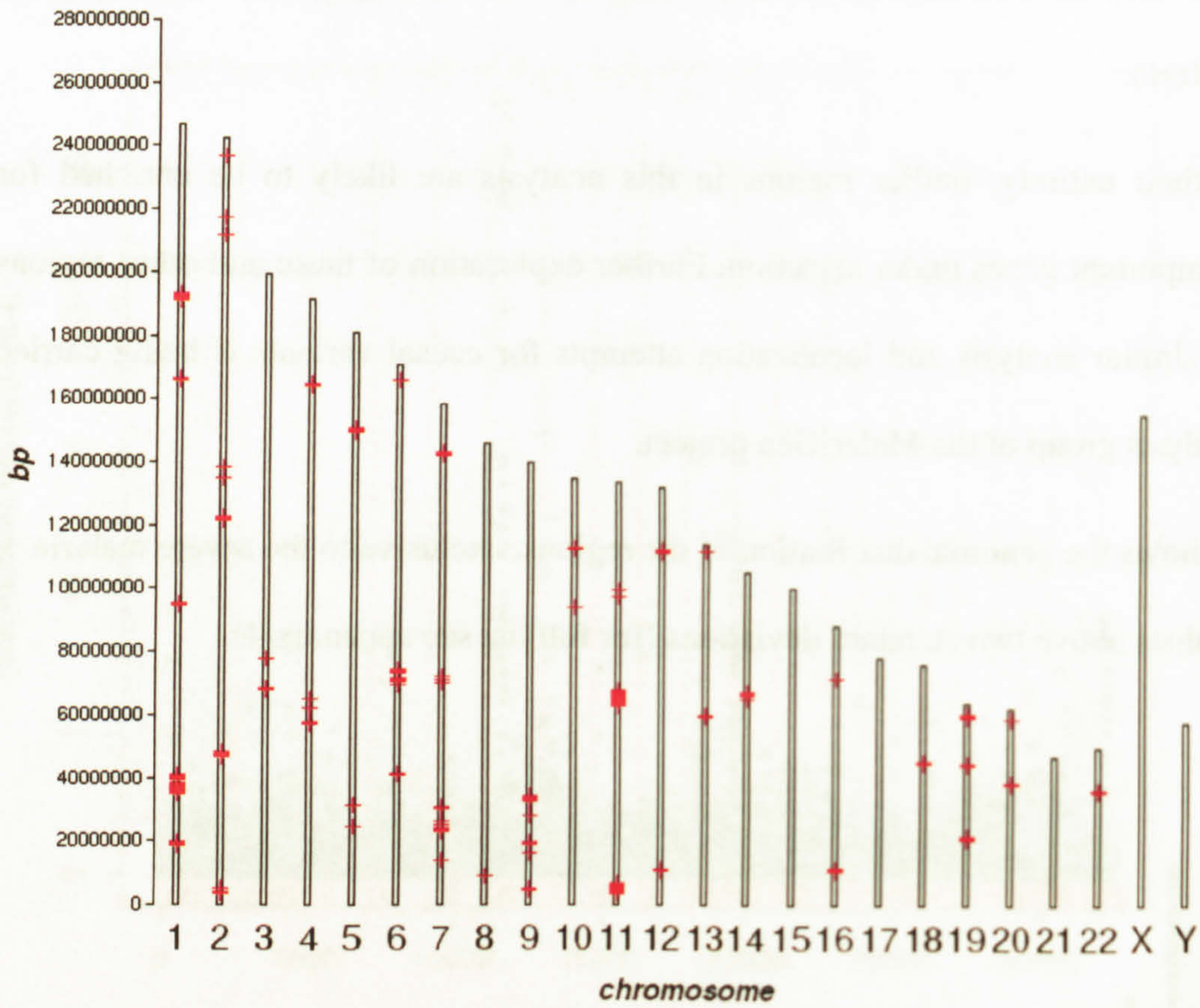


Figure 7.5.9: Genomic distribution of selection candidate regions exclusive to malaria controls (untransmitted haplotypes). Figure created using Chromosome Distribution Chart sub-module of WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/).

Figure 7.5.10 shows the genomic distribution of the regions, identified in both the cases and the controls to be outliers, with values above two standard deviations. These are regions that are potentially under positive selection, but not necessary malaria related (for full list see appendix 4).

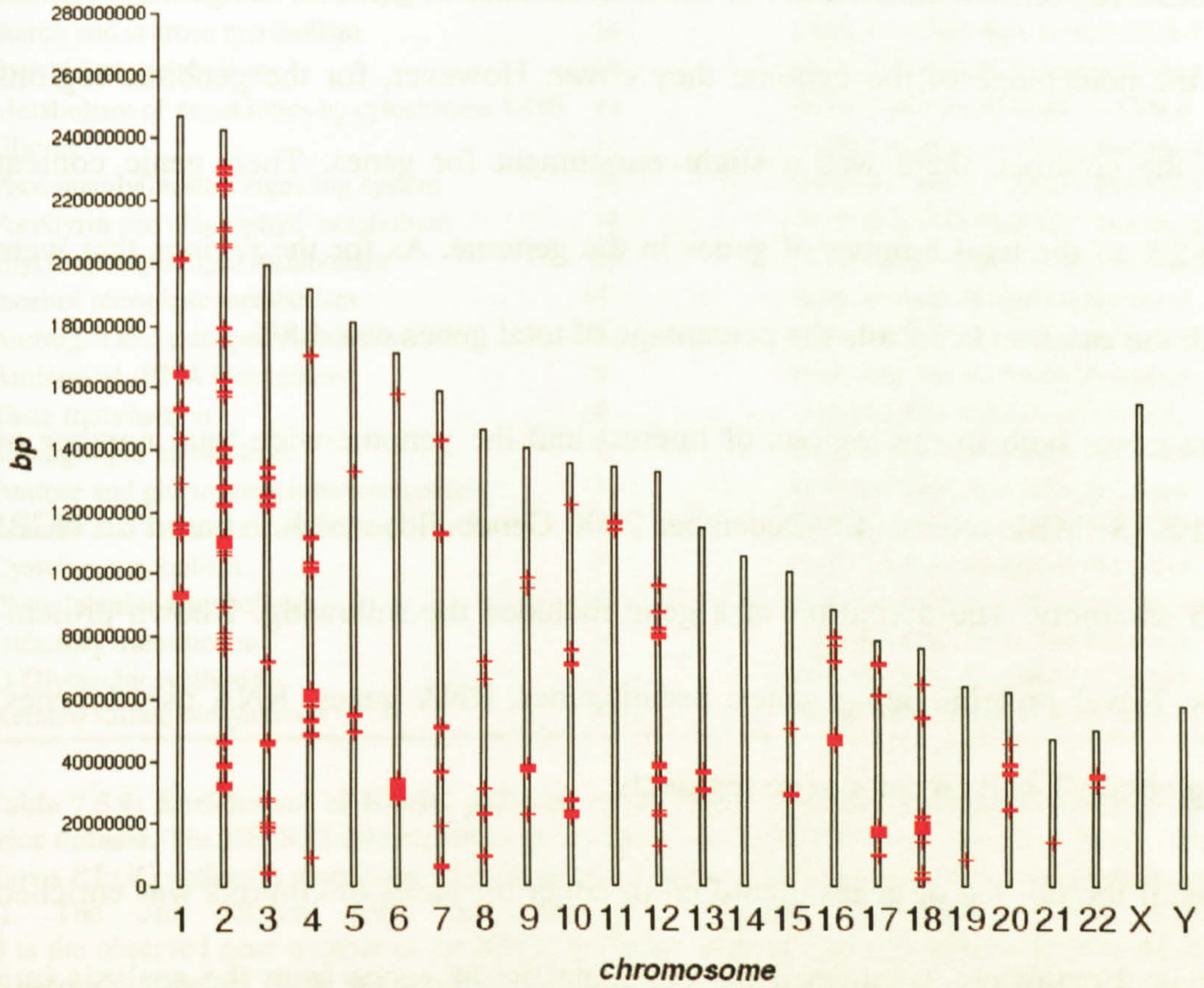


Figure 7.5.10: Chromosome distribution chart of candidate regions of recent adaptive evolution in both malaria cases and controls (positive in both transmitted and untransmitted chromosomes). Figure created using Chromosome Distribution Chart sub-module of WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/).

The outlier genomic regions that were exclusive to the malaria cases spanned 75 Mb in total, which corresponds to 2.3% of the genome. The total area of genomic regions exclusive to the malaria controls spanned a total of 70 Mb. This area is equivalent to 2.2% of the genome. For the regions that were shared between the cases and controls, they covered 138 Mb of the genome, equivalent to 4.2 %.

In the cases these regions contained 2.3% of the total number of genes in the genome, which agreed with the percentage of the genome they cover. However, for the genomic regions identified in the controls, there was a slight enrichment for genes. Their genic content constituted 3.2% of the total number of genes in the genome. As for the regions that were shared in both the cases and controls the percentage of total genes was 4.8%.

To define the genes both in the regions of interest and the genome-wide total number of genes, I used ENSEMBL release 42 (December 2006 Genebuild) which is based on NCBI 36, Oct 2005 assembly. The definition of a gene included the following: Known protein-coding genes, Novel protein-coding genes, pseudogenes, RNA genes, RNA pseudogenes, and immunoglobulin/T-cell receptor gene segments.

In order to see if the full list of genes identified in either the cases or controls was enriched for any biological pathways, I uploaded the complete list of genes from the analysis into ***WebGestalt***. Using the *KEGG Table and Maps* sub-module, ***WebGestalt*** organizes genes based on the KEGG biochemical pathways database (Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/). When analyzing the functional significance of the interesting gene set, all of the genes in human were used as the reference. In a KEGG Table (Table 7.5.4) the pathways associated with the gene set are shown, as well as the number of genes in each pathway, and the statistical parameters for the enrichment for the pathways.

KEGG pathway	Gene number	Enrichment
Cytokine-cytokine receptor interaction	34	O=34; E=13.8842; R=2.4488; P=1.33e-6
Focal adhesion	24	O=24; E=10.9008; R=2.2017; P=2.43e-4
Antigen processing and presentation	24	O=24; E=4.1308; R=5.81; P=7.67e-13
Regulation of actin cytoskeleton	22	O=22; E=11.3024; R=1.9465; P=2.21e-3
Calcium signaling pathway	20	O=20; E=9.5239; R=2.1; P=1.39e-3
Insulin signaling pathway	20	O=20; E=7.4585; R=2.6815; P=5.23e-5
Axon guidance	19	O=19; E=6.9421; R=2.7369; P=6.00e-5
Purine metabolism	19	O=19; E=8.3764; R=2.2683; P=7.13e-4
Natural killer cell mediated cytotoxicity	18	O=18; E=6.8274; R=2.6364; P=1.52e-4
Type I diabetes mellitus	16	O=16; E=2.2949; R=6.972; P=2.18e-10
Glycan structures - biosynthesis 1	16	O=16; E=5.6799; R=2.817; P=1.58e-4
Cell adhesion molecules (CAMs)	16	O=16; E=7.1142; R=2.249; P=1.98e-3
Starch and sucrose metabolism	16	O=16; E=3.7866; R=4.2254; P=7.17e-7
Pyrimidine metabolism	14	O=14; E=5.1062; R=2.7418; P=5.29e-4
Metabolism of xenobiotics by cytochrome P450	14	O=14; E=3.2703; R=4.281; P=2.97e-6
Glioma	11	O=11; E=3.4424; R=3.1954; P=5.46e-4
Phosphatidylinositol signaling system	10	O=10; E=3.9587; R=2.5261; P=5.81e-3
Porphyrin and chlorophyll metabolism	10	O=10; E=1.6064; R=6.2251; P=1.91e-6
Glycerophospholipid metabolism	10	O=10; E=3.5571; R=2.8113; P=2.64e-3
Inositol phosphate metabolism	9	O=9; E=2.6392; R=3.4101; P=1.06e-3
Androgen and estrogen metabolism	9	O=9; E=2.467; R=3.6482; P=6.31e-4
Aminoacyl-tRNA biosynthesis	9	O=9; E=1.7786; R=5.0602; P=4.21e-5
Taste transduction	8	O=8; E=2.467; R=3.2428; P=2.77e-3
Sphingolipid metabolism	8	O=8; E=2.0654; R=3.8733; P=8.25e-4
Pentose and glucuronate interconversions	7	O=7; E=0.8606; R=8.1339; P=8.65e-6
Alanine and aspartate metabolism	7	O=7; E=1.7786; R=3.9357; P=1.57e-3
Cysteine metabolism	7	O=7; E=1.2622; R=5.5459; P=1.61e-4
Phenylalanine metabolism	6	O=6; E=1.5491; R=3.8732; P=3.69e-3
Olfactory transduction	6	O=6; E=1.6064; R=3.7351; P=4.47e-3
O-Glycan biosynthesis	5	O=5; E=1.0327; R=4.8417; P=2.82e-3
Keratan sulfate biosynthesis	4	O=4; E=0.8606; R=4.6479; P=8.85e-3

Table 7.5.4: Enrichment of KEGG pathway in the list of candidate regions of selection in the Gambian trios dataset. The KEGG Table organizes genes based on the KEGG biochemical pathways. The KEGG Table shows KEGG pathways associated with the gene set (column 1), the number of genes in each pathway (column 2). The 3rd column gives the parameters for the enrichment of the KEGG pathway. O is the observed gene number in the KEGG pathway. E is the expected gene number in the KEGG pathway (Expected number of genes in a specific KEGG pathway for an interesting gene set=Total number of genes in the KEGG pathway for the reference set * Total number of genes in the interesting set / Total number of genes in the reference set). R is the ratio of enrichment for the KEGG pathway (R=O/E). P is the p value indicating the significance of enrichment calculated from Hypergeometric test. The KEGG pathways listed here are only those with at least 4 genes and with $p<0.01$.

7.6. Discussion

Outlier approaches, in which candidate selection genes are identified in the extreme tails of empirical distributions, have become a widely used strategy in genome-wide scans for selection (Akey, Zhang et al. 2002; Payseur, Cutter et al. 2002; Kayser, Brauer et al. 2003; Storz, Payseur et al. 2004; Voight, Kudaravalli et al. 2006).

In general, the simple outlier approach considered here is likely to result in an enriched set of genes that have been targets of positive selection. However, false discovery rate (FDR) can be high, depending upon parameters such as the strength of selection and the fraction of all loci that have been subject to selection. Unfortunately, these parameters are generally not known and are difficult to estimate.

In this analysis, I took into account variations in rates of recombination and number of markers, but I did not take into account variation in rates of mutation, and selection coefficients across loci, nor did I consider demographic perturbations that real populations are likely to experience. These factors are expected to increase variance and further complicate simple outlier approaches. There is no escaping the fact that evolutionary processes are inherently stochastic and extreme outlier values might arise under neutrality.

In this regard, the utility of simple outlier approaches may seem questionable. However, if the goal of a study, like it is in this case, is to identify a restricted set of candidate selection genes to study in more detail, then an outlier approach is a reasonable study design as long as one accepts that a substantial proportion of candidates may be false positives.

The strongest signal exclusive to the controls originated in the HBB region of chromosome 11 and covered a 3.4 Mb genomic area. This observation validated the ability of the method to identify genuine signals of selection. The classic examples of sickle cell anaemia and HbC represent some of the best examples of natural selection acting on the human genome. The

HBB signal was the second strongest signal genome wide. By far the strongest signal in both the cases and controls was that observed in the MHC region in chromosome 6.

There is a large body of evidence for the involvement of the MHC locus in malaria susceptibility. Piazza et al., were among the first to present evidence of the association between particular HLA variants and malaria in Sardinia, where they compared lowland areas where malaria occurred and highland areas (Piazza, Mayr et al. 1985). A case-control study in the Gambia indicated that the HLA class I antigen HLA-B53 and the HLA class II haplotypes DRB1*1302-DRB1*0501 both protect against severe malaria (Hill, Allsopp et al. 1991). In population studies, these genotypes accounted for as great a reduction in disease incidence as the sickle cell polymorphism, conferring 40% reduction in life-threatening complications of malaria in Gambian children (Hill, Bennett et al. 1992). In spite of the considerable literature on the subject, the results of this analysis indicate that the selective sweep observed in the MHC region is probably not related to malaria, at least in this Gambian sample set.

Few other significant signals - albeit less striking - were identified in the cases and controls. The regions highlighted by the analysis in malaria cases and controls, comprise a number of genes which may be classified as functional candidates because their products are operative in immune regulation or red blood cell metabolism, or other biological pathways suspected to play a part in malaria pathogenesis. While further pursuit of such signals is to be carried out by the MalariaGen analysis group, a preliminary analysis of the genic content and enrichment for biological pathways in regions identified as outliers, gave interesting results. The top three biological pathways enriched for in the full list of genes in the outlier regions were; cytokine-cytokine receptor interaction; focal adhesion; and antigen processing and presentation.

Cytokines induced by malaria products are a major determinant of disease progression. Upregulation by inflammatory cytokines of adhesion sites on endothelial cells invites susceptible circulating blood elements to attach to the inner wall of blood vessels (Michelson, Wencel-Drake et al. 1994).

Several investigators have posited that complex disease genes may be enriched for signatures of selection (Bamshad and Wooding 2003; Akey, Eberle et al. 2004), which can be regarded as an extension of the thrifty gene hypothesis proposed by Neel to explain the high prevalence of type II diabetes (Neel 1962). If this is in general true, then the genes that were found to possess evidence of selection in the malaria cases may be strong candidate disease genes.

A number of genome-wide scans for positive selection have recently been performed on the HapMap data (HapMap 2005; Sabeti, Varilly et al. 2007), which provide an important opportunity to compare results across studies. Forty-one regions out the195 autosomal regions identified in the HapMap samples using the iHS and LRH metrics (Sabeti, Varilly et al. 2007), were in the top fifth percentile of my analysis on the Gambian trios. Four out of the 26 autosomal genes with highly differentiated nonsynonymous SNPs described in Table 9 of The International HapMap Consortium (HapMap 2005) are among my candidate selection regions.

The considerable overlap of candidate selection genes with other genome-wide analyses engenders confidence in the method's predictions. However, it is important to confirm these results on independent data with analyses that test different predictions of neutrality, functionally characterize suspected targets of selection, and ultimately correlate adaptive genetic variation with phenotypic variation.

Although there is overlap between my results and previously described genome-wide scans for positive selection, there is also evidence for selection in genes not implicated in the above-described studies. This is to be expected for a number of reasons. For example, FDRs of outlier approaches are likely to be high. Furthermore, tests of neutrality generally have low statistical power.

7.7. Conclusion

In this chapter, I further developed the method introduced in chapter 6, and applied it to genome-wide data of severe malaria cases and controls from the Gambia. This method is built upon the premise that natural positive selection can be reflected in genetic variation patterns by the presence of unusually long haplotypes of relatively high frequency.

The analysis identified a number of interesting genomic regions that had unusually long haplotypes compared with the genome average. Although, this type of outlier approach employed here is bound to result in a high false discovery rate, it is likely that in their entirety, the outlier regions are enriched for biologically important genes undergoing selective sweeps. Certainly, the benchmark example of HbS has come up as one of strongest signals genome-wide and it was exclusively observed in the malaria controls. However, the strongest genome-wide signal originating from the MHC region was of equal magnitude in both the cases and controls, indicating its independence from malaria selective pressure.

The validity of this method is also supported by the overlap between regions identified by my analysis and those picked up by other genome-wide scans in other populations.

This analysis resulted in highlighting genomic regions enriched for interesting pathways with biological functions that might be implicated in malaria pathogenesis. The results

obtained from this study underline some of the regions in the genome where future detailed studies could be focused.

Chapter 8:

Summary and Discussion

There are many challenges facing the design, interpretation, and replication of genetic mapping studies of complex diseases. These challenges are much more manifest in African populations, due to the relatively complex population genetic patterns, and relative ethnic separation between many of the African populations with distinct ancestries.

In Africa there is also great variation in climate, diet, and exposure to infectious disease, which results in high levels of genetic and phenotypic variation. A better understanding of levels and patterns of variation in African genomes, together with phenotype data on susceptibility to disease, will be critical for identifying variants that play a role in susceptibility to a number of complex diseases and the development of more effective vaccines and other therapeutic treatments. No other point demonstrates the challenges facing association studies better than the fact that numerous association studies could not be replicated.

In general, these studies rely on background marker correlations to detect disease association. Therefore, understanding the structure of haplotypes in the human genome provides an important starting point for the study of complex traits. How patterns of LD vary between populations of different ethnic origin is highly topical, for two main reasons. First, variation in LD structure is integral to the problem of defining haplotype-tagging strategies that are transferable across different populations in disease association studies. Second, uncorrected population stratification may lead to false positives in association studies when there are systematic differences in the ancestry of cases and controls.

Through recurrent exposure to different pathogens, a number of genetic adaptations have evolved that provide resistance to infection in humans. Although the number of known candidate genes related to infectious disease has expanded, progress in the identification of genes that influence infectious disease susceptibility and/or resistance in diverse African populations has been slow. Studying the signatures of selection in African populations may provide a useful tool for gaining a clearer understanding of genetic and phenotypic adaptations in Africa.

During the course of this thesis, I gained insights into genomic signatures of recent positive selection and population differentiation in ethnically diverse African population groups. These insights can potentially be employed to better inform the design and interpretation of association studies in these groups.

Two ethnically distinct populations (The Hausa and Masalit) that inhabit neighbouring villages in eastern Sudan, were chosen for my study because each represent a typical African village with ethnically-homogenous, closely-related extended families. Furthermore, the two villages are located in an area endemic for malaria and visceral leishmaniasis (VL), and there is preliminary evidence of differential genetic susceptibility to these infectious diseases between the Hausa and Masalit. A population-based study was carried out to investigate host and parasite genetic factors that might underlie these differences in VL disease susceptibility.

At the start of this thesis, I set out to explore genetic diversity patterns in the Hausa and Masalit of Eastern Sudan, in an attempt to investigate if there was any genetic component to their apparent differential susceptibilities, in order to inform the design of future association studies to be carried out in these populations. As the 5q31 region is a susceptibility locus for parasitic infections including malaria and VL, I compared the haplotypic structure across the

region in the two populations. A dense concentration of immune genes is located in the human 5q31 region. The Th2 cytokine cluster in the human 5q31 region is known to be under coordinate regulation and several important enhancers and repressors have been defined at locations distinct from the genes themselves. I therefore examined long-range haplotype structure rather than limiting comparison to the variation at each individual gene locus. I studied a 656 kb segment of the 5q31 region containing 13 genes including *IL3*, *CSF2*, *IRF1*, *IL5*, *IL13* and *IL4*.

I genotyped 96 Hausa individuals and 96 Masalit individuals for 34 SNPs in the 5q31 region. These samples were mostly unrelated trios chosen to represent the population in each village. The process of choosing these trios involved constructing the whole village pedigree. From this it became apparent that there is a high degree of relatedness between individuals from different families within the same village. The whole village could be divided into several clusters where there are many strong ties between families. From a recent study carried out in the Masalit – the same group studied here- whole genome scan data showed there are only a limited number of Y chromosomal lineages in Salala village (Miller, Fadl et al. 2007). Although this setting – extended pedigrees with high degree of relatedness- might be ideal for some study designs like linkage studies and Family Based Association Testing (FBAT), it could present some challenges for others. For example it could potentially confound the results of case control studies, especially in founder populations that have grown rapidly and recently from a small size –as probably is the case here- where there would be an increased likelihood of sampling bias toward collecting relatives (Voight and Pritchard 2005). The above underlined the importance of careful consideration of schemes adopted when sampling from groups with high degrees of relatedness between their members.

The results of haplotype estimation in this dataset could have several roles in future disease-association studies to be carried out in these populations. For example, because the number of individuals sampled is likely to be sufficient for accurate estimation of haplotypes, using phased haplotypes with high probabilities from this small set of trios could help in phasing the haplotypes of a larger case control set of unrelated individuals from the same populations. Furthermore, the set of tagSNPs identified could potentially help with designing associations studies focused on the 5q31 region.

The markers I typed in the 5q31 region in the Hausa and Masalit were previously identified by members of the Kwiatkowski group, in the WTCHG laboratory in Oxford, as the most informative set (tagSNPs) in a sample from the Gambia in West Africa. The logic I drew on was that these markers being a tagSNP set from another African population were more likely to be useful in outlining the genetic variation pattern in the two Sudanese populations. In the past, the transferability of tagSNPs across populations, especially those from the same continental region, was suggested by several studies (Gu, Pakstis et al. 2007). Although Gonzalez-Neira et al. (Gonzalez-Neira, Ke et al. 2006) found Africa to be the most diverse region for the portability of tagSNPs from one population to another, nonetheless, they still found tagSNPs to be highly portable between African populations. However, those results were obtained in a gene-free region and may not be extended to other regions with different properties, like the 5q31 region.

Marker choice is likely to have had a major effect on different genetic variation patterns observed in the data. First of all, the amount of LD is likely to have been greatly affected by it. In both the Hausa and Masalit samples I observed little LD between markers. The average LD value was 0.05 with a variance of 0.01. LD values ranged from 1.8E-05 to 0.96 in the Hausa, and from 1.0E-05 to 1 in the Masalit. There was less LD and more diversity in the two Sudanese samples when compared with HapMap CEU sample. This is contrary to the

extensive LD expected in small semi-isolated populations due to bottle necks, small effective population size and inbreeding, especially when compared to a sample representing the whole population of Utah. The spacing and choice of markers from tagging SNPs which by definition have no or little LD between them might have made it more likely to get this result. A less likely alternative explanation, but one which could not have been dismissed with the available data, is that the low LD observed in the 5q31 region could be related to the important functionality of the region, being packed with genes involved in many aspects of immunity to a wide range of diseases. Fine-scale genetic map estimates from phase 2 HapMap data found genes involved in defence and immunity to have the highest recombination rates compared to genes of other functional classes (Frazer, Ballinger et al. 2007).

The second major effect of marker choice was that the sampled Hausa and Masalit were very similar in their minor allele frequencies, and all the tests that were run to try and differentiate (single marker Fst, haplotypic Fst, clustering methods like ARLEQUIN, **STRUCTURE** and genetic trees.) failed to identify the two populations as genetically distinct from each other, and to cluster individuals correctly. These observations went against the social, historical and linguistic evidence of their separation.

With the limited data generated in the 5q31 region, it was hard to distinguish whether the high degree of correlation in minor allele frequencies between the two Sudanese populations is an expected phenomenon resulting from the lack of resolution of the typed markers, a characteristic of the genomic area, or, alternatively, a reflection of real similarities due either to populations admixture, or convergent evolution due to balancing selection. Any of the above possibilities could be responsible for creating such a picture, but the most likely explanation is that this pattern is a consequence of the density, spacing and choice of

markers. It could very well be that the number and characteristics of typed markers, does not allow for enough resolution to distinguish these two populations from each other.

Although there were striking similarities in allele frequencies and amount of LD, and less than expected structuring by available genetic distance estimates; significant differences in haplotype composition were found to exist between the geographically contiguous Hausa and Masalit of Eastern Sudan. Marker choice and genomic processes such as mutation and recombination rates could affect the populations in the same manner, giving similar allele frequency patterns and overall quantity of LD, whereas the differences found in haplotypes, might be a reflection of the randomness of the sampling process, or demographic processes, such as expansions, founder effects and migrations. Previous studies of LD patterns in the human genome have shown that LD is sensitive to the demographic history of a population, like founder events, bottlenecks and isolation (Slatkin 1994; Service, DeYoung et al. 2006).

The processes by which SNPs have been selected by choosing high frequency markers from publicly available databases affect allele frequency spectrum more so than levels of LD observed in the data. While high frequency variants are more likely to be old and shared between population groups, consequently displaying little frequency differences between compared groups; these high frequency variants are more valuable in highlighting historical recombination events because of their higher resolution.

To explore this issue further, I tried to maximize the information content of markers that is used to tease out the genetic distinctness of the Hausa and Masalit, by comparing the LD patterns between these groups for this limited dataset. I also tested for the genetic differentiation among additional African (Gambians, YRI) and non-African (CEU) population groups, using the same approach to investigate its validity. I used similar sets of

polymorphic genetic markers (23-30 SNPs), typed in the same segment of the 5q31 region (about 650 kb).

Initially all the pair-wise r^2 values were calculated for all markers, within each group separately, using the Expectation Maximization (EM) algorithm. Afterwards the Spearman's rank correlation coefficient (rho) was calculated for the r^2 values between the two compared groups. Each r^2 value in the first group was paired to the corresponding r^2 of the same marker pair in the other group. For estimating the probability distribution and P-values, a series of bootstrap sampling was carried out, each time constructing two new groups from the pooled sample of individuals from the two populations together. Individuals were randomly selected from the pooled sample, ignoring their ethnicity assignment, to create two random groups of the same sizes as the real groups. Then for each of the two new random groups, pair-wise r^2 values were calculated, as well as the Spearman's rank correlation coefficient correlation, as done for the real groups. This process was repeated between (1000 to 50000 times). The P-value for obtaining the result of the real data is calculated from the distribution of the permutations' rho values, as the number of rho values equal or less than the real data rho value divided by the total number of permutations.

Although LD was found to be quantitatively very similar between the African population groups compared -The median and average of LD values, as well as their variance and range is comparable for population groups- the pattern of LD was shown to be different between them. Comparing the Hausa and Masalit samples; the Spearman's rank correlation coefficient (rho) was found to be (0.411878). When 10,000 permutations were carried out, correlation coefficients obtained from each of these permutations had a normal shaped distribution and from this distribution the P-value of observing the real data was found to be 0.016. Out of the six between-African population comparisons, four comparisons yielded significant results, when significance level was set to 0.05. Interestingly the pair that were

least correlated in their LD patterns were the Hausa and YRI which showed the highest degree of correlation between their minor allele frequencies. The minor allele frequency similarities could be due to their very close origin, as both of these populations are originally from Nigeria, but their LD pattern difference could be because the LD patterns reflect more recent demographic events which reflect ancestry rather than ethnicity.

When CEU and YRI were compared, at least 50,000 permutations were run before lower-than-real-data rho values were obtained by chance. When the HapMap CEU population was compared with the African groups, there was more than forty fold decrease in the P value (lowest P value 0.00002 compared to 0.0008). The genetic distance as reflected by rho and probability of genetic differentiation as reflected by the P-values; were found to be significantly more pronounced than the differences between African populations. This is probably due to the combined effects of more pronounced differences in allele frequencies as well as quantity and pattern differences in LD between the European and African samples.

Using this approach I managed, to a large extent, to tease out the distance between populations as predicted by their self specified ethnicities. I argue that any low correlation in pair-wise LD patterns between the study samples is due to the fact that the two samples come from two distinct ethnic groups with different demographies, rather than chance sampling effects resulting from sampling from the same population (the stochastic nature of sampling in a finite population). The permutation approach employed by this method to test the null hypothesis provided an inherent mechanism to estimate confidence in the results.

When comparing these results with those of other metrics of genetic distance estimation; very low values were found between groups of African populations when allele frequency-dependant metrics were applied. Furthermore, the approaches that utilized the full haplotypic information of all typed markers resulted in maximum values across all comparisons, which

probably represent an overestimation of the between-populations genetic distances. This indicates the unsuitability of these metric to analyse these data sets. These commonly used measures of population diversity or genetic distance consider either allele frequencies or haplotype frequencies. The allele frequency based methods, require large numbers of markers to be typed at unlinked loci, while the haplotypes based methods require a large number of sampled individuals from each group to accurately estimate diversity. So using methods based on allele frequency comparison may not be the most efficient approach in this setting. Not only does it not utilize the full information content of the data, but some methods recommend the exclusion of pairs of strongly linked loci that potentially bias the results.

One major advantage of using an LD based method to highlight between-populations genetic differences is that LD is highly relevant in disease association studies context. Discerning populations' genetic differentiation by utilizing LD lends itself to the analysis of case control association studies by being potentially more sensitive in highlighting population structure that might be undetected by looking at allele frequencies alone, especially when data is limited. The pattern and extent of LD determines the feasibility and design of association studies when haplotypes are used to test associations or when untyped SNPs are imputed. Testing for LD pattern homogeneity between groups making up the sample goes a longer way towards minimizing type 1 error than relying on allele frequency information alone. Considering LD in association study design is much more relevant in African populations. Some evidence has suggested variance in levels and patterns of LD among subpopulations in Africa. Tishkoff and colleagues (Tishkoff, Dietzsch et al. 1996) noted that African populations have divergent patterns of LD; specifically, alleles that were in positive association in one population were in negative association in another. Additionally, a resequencing analysis of the IL-13 gene in 126 geographically diverse Africans identified

divergent patterns of LD across West and East African populations (Tarazona-Santos and Tishkoff 2005). These observations suggest that not all African populations are characterized by a single discrete pattern of LD and each may have distinct haplotype block structures.

It might also be useful to quantify the degree to which LD relationships will hold when attempting to transfer and judge the coverage of tagSNPs when using data from one population like the HapMap samples, in designing studies and guiding analysis in other populations. Recently, some studies have revealed significant variation in the underlying haplotype structure in spite of the observed conservation of tagSNP patterns across global populations (Gu, Pakstis et al. 2007). This might indicate that even in cases where the coverage of tagSNPs appears to be preserved across populations, caution still needs to be exercised because the hidden genetic variants tagged by any particular tagSNP might not be the same in different populations.

It is unlikely that a single best method can be recommended for the estimation of genetic differentiation, but this approach is a useful addition to existing methods for estimating genetic distance, and it shows promise in computing the genetic distance from the correlated structure of genomic variation. This proposed approach could have a practical significance in analyzing similar datasets, with the potential for future applications in deciphering stratification in population samples and case control studies.

Another major area of population genetics which this thesis encompasses is investigating how the signals of positive selection are reflected in the genetic polymorphism patterns. Identifying positive selection signals in a genomic region of interest offers a much needed insight into the search for disease modifying variants.

The need for better characterization of these kinds of signals in the Hausa and Masalit arose from the ambiguous results previously obtained when analyzing the 5q31 region in the two

groups. There was no clear selection signal in the 5q31 in the Hausa. The Masalit data had a small signal that was very close to the background noise in the region. In order to verify the significance of any signals detected, a comparison with other regions of the genome had to be made in the two populations. Close examination of another area of the genomes of these two populations where natural selection is known to have played a part in shaping its diversity and where the functional variant is known (the β-globin region harbouring the sickle variant); allowed an easier interpretation of the genetic variation patterns associated with positive selective pressures in the Hausa and Masalit.

The β-globin region being the classical example of a locus under positive selection, offers the perfect opportunity for bench marking the other less defined positive selection signals in the genome. This approach makes available prior knowledge of the functional variant and its attributes, i.e.: position, frequency, LD, and haplotypic relationship with other markers in the region. Applying knowledge gained from, and looking for patterns recognized in the β-globin to the 5q31 region, helped to better interpret its genetic variation results.

The malaria hypothesis maintains that during prehistory, on average, people without the sickle gene died of malaria at a high frequency. On the other hand, people with two genes for sickle haemoglobin died of sickle cell disease. In contrast, the heterozygotes (sickle trait) were more resistant to malaria than normal individuals and yet suffered none of the ill-effects of sickle cell disease. This selection for heterozygotes is termed "balanced polymorphism". Support for this concept comes from epidemiological studies in malaria-endemic regions of Africa. A recent study found that the state of having one sickle cell allele was associated with protection against mild clinical malaria (50%), hospital admission for malaria (75%) and severe malaria (90%). The parasite densities during clinical attacks in children with HbAS were also found to be lower than HbAA children (Williams, Mwangi et al. 2005).

The Hausa group displayed a very distinct selection signal in the β-globin region. But the detection of the signal was conditional on including the functional marker itself –the HbS polymorphism- in the analysis. The observation that sickle cell haplotypes, in stark contrast to others, were the highest in frequency across the 400 kb region, prompted an in-depth look at how far out these haplotypes extend, because such phenomenon could be a surrogate for positive selection signals in the genome, I set out to better characterize it by attempting to answer the following two questions: Firstly, is the high frequency extended HbS haplotype exclusive to the Sudanese populations or can it be discernable in another African population. Secondly, are there similar instances in other genomic regions, and if so, is there any supporting evidence of them being candidates of positive selection.

I used publicly available genome-wide genotyping data from the Yoruba (YRI) samples that were typed in the HapMap project. To make a meaningful comparison with the Sudanese data, I typed the 90 YRI samples for SNPs typed in the Sudanese samples that had not been typed in the HapMap project. Publicly available HapMap data was also used for thirty trios which were collected from U.S. residents with Northern and Western European ancestry (CEU).

I developed a Perl script to run on a UNIX platform in order to look for instances of unusually extended high-frequency haplotypes in the genome. The Perl code was used to scan the human genome employing an overlapping window approach. The script looks at all the haplotypes within a predefined window. All chromosomes of both the YRI and CEU were scanned using window size prefixed to 360 markers. The window slides across haplotypes supplied (phased HapMap data) by shifting the window position along the length of each chromosome. Firstly, I used window shift of 180 markers, so as to make windows overlap by half their sizes. Then I carried out another genome scan with the much smaller window shift of 1 marker.

This script also calculates the average recombination rate for each window position using a file of estimated recombination rates downloaded from HapMap.

After acquiring the data for all the windows across each of the chromosomes, the frequency of the highest identical haplotype in a window would be plotted against the genetic distance value for that particular window. This would create a chromosome-wide distribution amenable to an outlier analysis of all windows across a chromosome.

In the HBB region in the YRI, the highest haplotype frequency was that of an HbS haplotype. When compared with the Sudanese sample data, this observation was clearer in the YRI sample probably due to the higher marker resolution (165 markers in YRI as opposed to 26 markers typed in the same area in the Sudanese sample).

This high frequency haplotype was maintained for 1.2 Mb around the HbS allele, spanning several recombination hot spots before declining very rapidly to become indistinguishable from others in the same region. Also noted was the fact that over distances less than 200 kb the HbS haplotypes were grouped together with other HbA haplotypes because they were indistinct at the analysed marker density. The most likely explanation for this phenomenon is that in the YRI HBB region, malaria selection pressure acting on the sickle-cell variant helped maintain identical HbS haplotypes at this high frequency. This effect was equal on both sides of the HbS polymorphism (600 kb) regardless of the number, position and intensity of recombinational hotspots on each side. Therefore the effect of selection on haplotype frequencies does not seem to be correlated with the fine scale recombination rate but rather is tuned by the overall recombination rate in the region.

To determine whether this observation could be utilized as a method for identifying genomic regions under positive natural selection, it was important to quantitatively determine how

significant this finding is on a larger scale, when measured against the whole of chromosome 11 and the rest of the genome. Phased haplotypic data from HapMap phase1 was analysed for chromosome 11 using a sliding window approach In chromosome 11, the region that showed unusually high frequency long haplotype with the extended-high-frequency-haplotype analysis was the same region that demonstrated clustering of signals using other established haplotype-based methods for detecting selection like the LRH and haplosimilarity.

Phased haplotypic data from HapMap phase1 for both the YRI and CEU populations were analysed for each chromosome at a time, using the same sliding window approach. In total there were 55 regions that were picked up from the analysis. Twenty three regions in the YRI and 32 in the CEU. The average size of region was 2.78 Mb in YRI and 2.64 Mb in CEU samples. From the total 55 regions there were 8 regions shared between YRI and CEU and 39 regions exclusive to one or the other population. This is an interesting result suggesting that local adaptation has played an important role in recent human evolutionary history.

The total number of genes in the regions that stood out from YRI and CEU whole genome scan as possible candidates of positive selection, was 691 genes in a total area of 124002522 bp with average gene density of one gene every 179454 bp. Out of these 691 genes identified, 77 genes (10%) were genes involved in immunity. When this is compared with the 770 genes involved in immunity out of the 33524 total genes in the human genome (about 2%), it becomes clear that there is a higher preponderance of genes involved in immuntiy in the outlier regions identified by the extended-high-frequency-haplotype genomic scan. and is highly suggestive of them being selectively important. At least 18 of the 47 regions in both YRI and CEU had a previously reported evidence of being under natural selection pressure or had a positive signal in association studies.

The HBB region had the highest signal in the YRI genome, which is hardly surprising given the fact that the whole scan was optimized on this signal. In the CEU genome the most remarkable signal mapped to the 2q21.3 region within which the LCT gene is present. The LCT gene was previously found in Northern European populations to have very high frequencies of the lactase persistence allele (LCT*P) (Hollox, Poulter et al. 2001), which allows digestion of fresh milk throughout adulthood. It is widely accepted that strong selection has driven LCT*P to high frequency in Northern Europeans, beginning sometime after the domestication of animals approximately 9,000 years ago (Hollox, Poulter et al. 2001; Bersaglieri, Sabeti et al. 2004).

This analysis helped identify and describe the genomic scale over which selective sweeps could have an effect. With the extended-high-frequency-haplotype method, detecting positive selection in genomic regions could be achieved without regard to whether the causal variant was typed or not. Consequently, there is less emphasis on marker choice, density and spacing unlike other methods (like LRH and haplosimilarity) which rely on the ability of a marker to tag the causal SNP by being in high LD with it and thus making marker choice and density of essential importance. Using data for all SNPs in a genomic region in the order of a megabase makes this method robust to marker choice and density variation when compared to the above mentioned methods. The consistency in finding the high frequency extended haplotype in the face of variable marker density, and chance element in choice and ascertainment of typed markers, gives this method an advantage by decreasing the rate of false negatives when looking for signals of positive selection.

This property may make this method useful for genome wide case control studies on a large number of individuals with a modest marker coverage that will not necessary tag all the untyped markers, a thing which is logistically difficult to achieve either due to limitations in resources, technology or an over-fitting problem in marker choice which in most instances

rely on an imperfectly transferable SNP-tagging sets between populations and studies. This method will probably capture the same haplotypic diversity with less marker density. Furthermore, it is likely that the difference in frequency between selected and other neutral haplotypes in the same region will become more prominent with the larger number of individuals typed. When I reduced the density of the markers in the 1.1 Mb region around HbS to half that in phase 1 of the HapMap, by taking every other marker's genotypes out and then re-phasing the genotypic data. The HbS haplotypes were still distinguished from others by their high frequency over that distance. In a recent study (Conrad, Jakobsson et al. 2006) it was shown that the bigger the window considered, the more powered the genotyped SNPs to capture the haplotypic heterozygosity in the area as measured by microsatellites.

The most important insight from my analyses is highlighting the scale over which signals of selection are most effectively detected, and giving other methods of looking for natural selection context by considering all the members of a cluster of signals in a genomic region to be telling the same story. Using the extended-high-frequency-haplotype method is a simple and quick way to highlight a particular genomic region as a candidate of natural positive selection, as well as defining the boundaries of that region for further analysis. This type analysis will probably not be as informative without defining the unit size by the initial scan.

The challenge of this method stems from its source of strength. It becomes more of a challenge to pinpoint the functional variant, the larger the genomic area over which the search has to be conducted. Several markers in the regions identified by the extended high frequency haplotype method are expected to have unusually extensive LD values because of their close correlation with the high frequency selected haplotype.

Moreover, using the correlation between haplotype frequencies and recombination rate as a test to look for selective sweeps will miss those regions with no or very little recombination rates if they were acted on by positive selection. As it stands now this analysis is a conservative way to scan for selection, in the sense that it would only have power to pick up areas with incomplete selective sweeps which are relatively recent and did not yet reach fixation, due to the underestimation of recombination rates in regions with complete sweeps.

In the YRI I fixed the window size to 360 markers, which roughly corresponded to 1Mb. I chose this size because it is optimized to selective sweeps of a similar or greater magnitude to that observed for the HbS in YRI. The choice of window size, that would be optimal for identifying genomic regions under selection, warrants some consideration for different datasets. At very small window sizes, most haplotypes will be of a high frequency which makes them indistinct from the selected haplotype. At the other extreme of very big windows, all haplotypes would be distinct from each other leading to a failure to pick the selected haplotype. Between these two extremes of distribution uniformities, all the possible signals with different effect sizes could potentially be identified by running the analysis with different window sizes.

The extended-high-frequency-haplotype method developed in the HapMap data showed promising results, suggesting its utility in highlighting genomic regions that might be candidates of positive selection. Applying it to real-life genome-wide data of Gambian children with severe malaria and their parents as part of the Malaria Genomic Epidemiological Network (MalariaGEN) project, presented the opportunity to further develop and validate the method, and to identify genomic regions where positive selection might have played a role. Additionally, it was an opportunity to gather disease-specific inferences that might aid the search for malaria resistance/susceptibility genetic variants. 2632 chromosomes were analysed for 585,350 SNPs (total number of genotypes

770,320,600). Overlapping windows of 1cM size and 0.1cM shift were run across the 22 autosomes. Data from all chromosomes was then combined to carry out a genome-wide statistical assessment where the upper 2.3% of the data points were highlighted and further explored for their genic content. This analysis was carried out separately for the malaria cases and controls.

Each chromosome was analysed twice; once with the haplotypes that were transmitted from parents to children (representing the severe malaria cases), and a second time with the haplotypes that were not transmitted from the parents (representing the controls). Therefore, the analysis carried out in the cases can potentially highlight genomic regions where genetic variants might be involved in malaria susceptibility. On the other hand, the analysis carried out in the controls could potentially identify genomic regions where malaria protective polymorphisms reside. Each chromosome was done independently then the data was pooled for all transmitted chromosomes and statistical significance determined. The same was done for the controls on the parental haplotypes not transmitted to the malaria cases.

I was particularly interested in the HbS locus and the MHC regions, to see if these two genomic regions, which have extensive literature supporting their involvement in malaria susceptibility, would come up as significant in this analysis.

The classic examples of sickle cell anaemia and HbC represent some of the best examples of natural selection acting on the human genome. Therefore, the HBB region in chromosome 11 is the bench mark example of malaria-specific selection. It was of great interest to see whether this method would pick up a selection signal in the HBB region in the controls. When looking at the results from chromosome 11, there was a very strong and clear signal in the HBB region with the analysis that was carried out in the controls.

The signal originating from the HBB region constituted the highest signal in chromosome 11. It was the second strongest genome wide after the MHC signal, with values above the 13th standard deviation. It reflected the presence of an unusually long and high frequency haplotype in the HBB region. This haplotype which carries the HbS allele maintained a relatively high frequency over many overlapping windows where its frequency was outside the genome-wide 95th percentile. These windows covered a 3.4 Mb genomic area. The maximum frequency of 66 copies was in the 0.5 Mb (0.98 cM) around the HbS position (11:4795740_11:5309382). Furthermore, the selection signal at the HBB region disappeared completely when analyzing the haplotypes transmitted to the cases, which makes a strong argument that the pattern observed is the result of selection on the HbS allele in the Gambian controls. This observation validated the ability of the method to identify genuine signals of selection.

Where the MHC and other immune genes are located on chromosome 6, the selection signal was observed in both the transmitted haplotypes (cases) and untransmitted haplotypes (controls). The same haplotype was responsible for creating this signal in the both the cases and controls. The strength of this signal was greater than that observed in the HBB region in the malaria controls. By far the strongest signal in both the cases and controls was that observed in the MHC region on chromosome 6. Values were well above 25 standard deviations of the genome wide average. Due to the many immunologically important genes in this region, it is not implausible that the signal observed might be due to the effects of a very strong natural positive selection force acting on the region, but not necessary due to malaria.

The importance of genes regulating immune responses to malaria was demonstrated by the finding of HLA associations with resistance to severe malaria (Hill, Allsopp et al. 1991). Polymorphism in the promoter of another MHC gene, tumour necrosis factor TNF, was

found to affect the risk of cerebral malaria (McGuire, Hill et al. 1994). However it has been surprisingly difficult to detect an influence of HLA and other major histocompatibility complex genes on the magnitude of immune responses to malarial antigens in field studies. In general, cellular immune responses to malaria antigens show marked heterogeneity in specificity, type and magnitude; the relative importance of MHC polymorphism and other genetic factors in accounting for this heterogeneity has been unclear (Hill, Jepson et al. 1997).

There is a large body of evidence for the involvement of the MHC locus in malaria susceptibility. Piazza et al., were among the first to present evidence of the association between particular HLA variants and malaria in Sardinia, where they compared lowland areas where malaria occurred and highland areas (Piazza, Mayr et al. 1985). A case-control study in the Gambia indicated that the HLA class I antigen HLA-B53 and the HLA class II haplotypes DRB1*1302-DRB1*0501 both protect against severe malaria (Hill, Allsopp et al. 1991). In population studies, these genotypes accounted for as great a reduction in disease incidence as the sickle cell polymorphism, conferring 40% reduction in life-threatening complications of malaria in Gambian children (Hill, Bennett et al. 1992). In spite of the considerable literature on the subject, the results of this analysis indicate that the selective sweep observed in the MHC region is probably not related to malaria, at least in this Gambian sample.

A number of additional regions with evidence of unusually long high frequency haplotypes were identified in the cases and controls. Although their statistical significance might be weaker than regions mentioned above, but still they might be plausible candidates for positive selection. Experiences in other infectious diseases have shown that weaker signals may as well lead to the identification of relevant genetic variants (Ogura, Bonen et al. 2001).

Certainly in their entirety, outlier regions in this analysis are likely to be enriched for biologically important genes under selection. A preliminary analysis of the functional significance of the full list of genes identified in the outlier regions of the cases and controls highlighted certain biological pathways as being over-represented. The top three enriched biological pathways were; cytokine-cytokine receptor interaction; focal adhesion; and antigen processing and presentation. Cytokines induced by malaria products are a major determinant of disease progression. Upregulation by inflammatory cytokines of adhesion sites on endothelial cells invites susceptible circulating blood elements to attach to the inner wall of blood vessels (Michelson, Wencel-Drake et al. 1994).

A number of genome-wide scans of positive selection have recently been performed on the HapMap data (HapMap 2005; Sabeti, Varilly et al. 2007), which provide an important opportunity to compare results across studies. Fourty-one regions out the195 autosomal regions identified in the HapMap samples using the iHS and LRH metrics (Sabeti, Varilly et al. 2007), were in the top fifth percentile of my analysis on the Gambian trios. Four out of the 26 autosomal genes with highly differentiated nonsynonymous SNPs described in Table 9 of The International HapMap Consortium (HapMap 2005) are among my candidate selection regions.

The considerable overlap of candidate selection genes with other genome-wide analyses engenders confidence in the method's predictions. However, it is important to confirm these results on independent data with analyses that test different predictions of neutrality, functionally characterize suspected targets of selection, and ultimately correlate adaptive genetic variation with phenotypic variation. Although there is overlap between my results and these previously described genome-wide scans for positive selection, there is also evidence for selection in genes not implicated in the above-described studies. This is to be

expected for a number of reasons. For example, tests of neutrality generally have low statistical power. Furthermore, the FDRs of outlier approaches are likely to be high.

Outlier approaches, in which candidate selection genes are identified in the extreme tails of empirical distributions, have become a widely used strategy in genome-wide scans for selection (Akey, Zhang et al. 2002; Payseur, Cutter et al. 2002; Kayser, Brauer et al. 2003; Storz, Payseur et al. 2004; Voight, Kudaravalli et al. 2006). In general, the simple outlier approach considered here is likely to result in an enriched set of genes that have been targets of positive selection. However, FDRs can be high, depending upon parameters such as the strength of selection and the fraction of all loci that have been subject to selection. Unfortunately, these parameters are generally not known and are difficult to estimate.

In this analysis, I took into account variations in rates of recombination and number of markers, but I did not take into account variation in rates of mutation, and selection coefficients across loci, nor did I consider demographic perturbations that real populations are likely to experience. These factors are expected to increase variance and further complicate simple outlier approaches. There is no escaping the fact that evolutionary processes are inherently stochastic and extreme outlier values might arise under neutrality. In this regard, the utility of simple outlier approaches may seem questionable. However, if the goal of a study, like it is in this case, is to identify a restricted set of candidate selection genes to study in more detail, then an outlier approach is a reasonable study design as long as one accepts that a substantial proportion of candidates may be false positives.

Analyzing MalariaGen Gambian dataset for extended high-frequency haplotypes helped highlight a number of genomic regions which might harbour genes or biological pathways suspected to play a part in malaria pathogenesis. This work helped inform a much broader analysis - to further explore, improve and expand the search for such signals - that is being

carried out by the MalariaGen analysis group, the results of which are to be published by MalariaGen consortium in the future.

To summarize, this thesis has been about exploring two important aspects of population genetics that are of topical importance in the design and interpretation of association studies. These insights, although driven from and focused on the populations of this study, are not limited to them. Promising results have been obtained that might with further future work lead to improving the determination of sample substructure by considering the LD relationship between markers. More clues of disease susceptibility loci could potentially be gained from using the extended-high-frequency-haplotype method to look for signals of natural selection in the genome. This analysis does not require the typing of the causal variant. It also helps highlight the scale over which signals of selection are most effectively detected, and determine the boundaries of the region on which further more detailed search is to be conducted.

Although interesting regions worthy of future pursuit were identified in relation to malaria susceptibility, the method could be as useful in other disease studies. Certainly if time would have allowed I would have liked to further develop and test the robustness of the method for determining population differentiation, as well as the method for detecting natural positive selection in the genome by looking at high frequency extended haplotypes. By structuring them as formal methods with an online user interface, they would have been available to the MalariaGen consortium and the wider scientific community. Also I might have been involved in exploring the regions that came up as interesting in the analysis of MalariaGen Gambian population, with the objective of narrowing the search down to a few biologically relevant genes that might influence malaria susceptibility and might be amenable to functional studies.

REFERENCES

(2000). "Severe falciparum malaria. World Health Organization, Communicable Diseases Cluster." Trans R Soc Trop Med Hyg **94 Suppl 1**: S1-90.

Abecasis, G. R., D. Ghosh, et al. (2005). "Linkage disequilibrium: ancient history drives the new genetics." Hum Hered **59**(2): 118-24.

Agarwal, A., A. Guindo, et al. (2000). "Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S." Blood **96**(7): 2358-63.

Akey, J. M., M. A. Eberle, et al. (2004). "Population history and natural selection shape patterns of genetic variation in 132 genes." PLoS Biol **2**(10): e286.

Akey, J. M., G. Zhang, et al. (2002). "Interrogating a high-density SNP map for signatures of natural selection." Genome Res **12**(12): 1805-14.

Akey, J. M., K. Zhang, et al. (2003). "The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium." Mol Biol Evol **20**(2): 232-42.

Allen, S. J., A. O'Donnell, et al. (1996). "Severe malaria in children in Papua New Guinea." Qjm **89**(10): 779-88.

Altshuler, D., V. J. Pollara, et al. (2000). "An SNP map of the human genome generated by reduced representation shotgun sequencing." Nature **407**(6803): 513-6.

Anastasi, J. (1984). "Hemoglobin S-mediated membrane oxidant injury: protection from malaria and pathology in sickle cell disease." Med Hypotheses **14**(3): 311-20.

Andolfatto, P. (2001). "Adaptive hitchhiking effects on genome variability." Curr Opin Genet Dev **11**(6): 635-41.

Andolfatto, P., J. D. Wall, et al. (1999). "Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster." Genetics **153**(3): 1297-311.

Angius, A., D. Bebbere, et al. (2002). "Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations." Hum Genet **111**(1): 9-15.

Angius, A., F. C. Hyland, et al. (2008). "Patterns of linkage disequilibrium between SNPs in a Sardinian population isolate and the selection of markers for association studies." Hum Hered **65**(1): 9-22.

Ardlie, K., S. N. Liu-Cordero, et al. (2001). "Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion." Am J Hum Genet **69**(3): 582-9.

Ardlie, K. G., L. Kruglyak, et al. (2002). "Patterns of linkage disequilibrium in the human genome." Nat Rev Genet **3**(4): 299-309.

Bamshad, M. and S. P. Wooding (2003). "Signatures of natural selection in the human genome." Nat Rev Genet **4**(2): 99-111.

Bamshad, M. J., S. Wooding, et al. (2003). "Human population genetic structure and inference of group membership." Am J Hum Genet **72**(3): 578-89.

Bersaglieri, T., P. C. Sabeti, et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene." Am J Hum Genet **74**(6): 1111-20.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-48.

Bolad, A., S. E. Farouk, et al. (2005). "Distinct interethnic differences in immunoglobulin G class/subclass and immunoglobulin M antibody responses to malaria antigens but not in immunoglobulin G responses to nonmalarial antigens in sympatric tribes living in West Africa." Scand J Immunol **61**(4): 380-6.

Borst, P., W. Bitter, et al. (1995). "Antigenic variation in malaria." Cell **82**(1): 1-4.
Braverman, J. M., R. R. Hudson, et al. (1995). "The hitchhiking effect on the site frequency spectrum of DNA polymorphisms." Genetics **140**(2): 783-96.
Breman, J. G., M. S. Alilio, et al. (2004). "Conquering the intolerable burden of malaria: what's new, what's needed: a summary." Am J Trop Med Hyg **71**(2 Suppl): 1-15.
Bryceson, A. D., A. F. Fleming, et al. (1976). "Splenomegaly in Northern Nigeria." Acta Trop **33**(3): 185-214.
Bustamante, C. D., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-7.
Cann, R. L., M. Stoneking, et al. (1987). "Mitochondrial DNA and human evolution." Nature **325**(6099): 31-6.
Cardon, L. R. and G. R. Abecasis (2003). "Using haplotype blocks to map human complex trait loci." Trends Genet **19**(3): 135-40.
Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases." Nat Rev Genet **2**(2): 91-9.
Cargill, M., D. Altshuler, et al. (1999). "Characterization of single-nucleotide polymorphisms in coding regions of human genes." Nat Genet **22**(3): 231-8.
Carlson, J., G. B. Nash, et al. (1994). "Natural protection against severe Plasmodium falciparum malaria due to impaired rosette formation." Blood **84**(11): 3909-14.
Cavalli-Sforza, L., P. Menozzi, et al. (1994). The history and geography of human genes. Princeton, NJ, Princeton University Press.
Cavalli-Sforza, L. L. and M. W. Feldman (2003). "The application of molecular genetic approaches to the study of human evolution." Nat Genet **33 Suppl**: 266-75.
Chakravarti, A., K. H. Buetow, et al. (1984). "Nonuniform recombination within the human beta-globin gene cluster." Am J Hum Genet **36**(6): 1239-58.
Charlesworth, B., M. T. Morgan, et al. (1993). "The effect of deleterious mutations on neutral molecular variation." Genetics **134**(4): 1289-303.
Collins-Schramm, H. E., C. M. Phillips, et al. (2002). "Ethnic-difference markers for use in mapping by admixture linkage disequilibrium." Am J Hum Genet **70**(3): 737-50.
Collins, W. E. and G. M. Jeffery (1999). "A retrospective examination of the patterns of recrudescence in patients infected with Plasmodium falciparum." Am J Trop Med Hyg **61**(1 Suppl): 44-8.
Coluzzi, M., A. Sabatini, et al. (1979). "Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex." Trans R Soc Trop Med Hyg **73**(5): 483-97.
Conrad, D. F., M. Jakobsson, et al. (2006). "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome." Nat Genet **38**(11): 1251-60.
Cooke, B. M., N. Mohandas, et al. (2004). "Malaria and the red blood cell membrane." Semin Hematol **41**(2): 173-88.
Cot, M. and P. Deloron (2003). "Malaria prevention strategies." Br Med Bull **67**: 137-48.
Crawford, D. C., D. T. Akey, et al. (2005). "The patterns of natural variation in human genes." Annu Rev Genomics Hum Genet **6**: 287-312.
Crawford, D. C., T. Bhangale, et al. (2004). "Evidence for substantial fine-scale variation in recombination rates across the human genome." Nat Genet **36**(7): 700-6.
Currat, M., G. Trabuchet, et al. (2002). "Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation." Am J Hum Genet **70**(1): 207-23.
Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-32.

Day, N. P., T. T. Hien, et al. (1999). "The prognostic and pathophysiologic role of pro- and antiinflammatory cytokines in severe malaria." J Infect Dis 180(4): 1288-97.
Day, N. P., N. H. Phu, et al. (2000). "The pathophysiologic and prognostic significance of acidosis in severe adult malaria." Crit Care Med 28(6): 1833-40.
De La Vega, F. M., D. Dailey, et al. (2002). "New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies." Biotechniques Suppl: 48-50, 52, 54.
De La Vega, F. M., H. Isaac, et al. (2005). "The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern." Genome Res 15(4): 454-62.
Depaulis, F. and M. Veuille (1998). "Neutrality tests based on the distribution of haplotypes under an infinite-site model." Mol Biol Evol 15(12): 1788-90.
Diamond, J. (2002). "Evolution, consequences and future of plant and animal domestication." Nature 418(6898): 700-7.
Dolo, A., D. Modiano, et al. (2005). "Difference in susceptibility to malaria between two sympatric ethnic groups in Mali." Am J Trop Med Hyg 72(3): 243-8.
Eberle, M. A., P. C. Ng, et al. (2007). "Power to detect risk alleles using genome-wide tag SNP panels." PLoS Genet 3(10): 1827-37.
Emahazion, T., L. Feuk, et al. (2001). "SNP association studies in Alzheimer's disease highlight problems for complex disease analysis." Trends Genet 17(7): 407-13.
Evans, D. M. and L. R. Cardon (2005). "A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations." Am J Hum Genet 76(4): 681-7.
Excoffier, L. (2002). "Human demographic history: refining the recent African origin model." Curr Opin Genet Dev 12(6): 675-82.
Fallin, D. and N. J. Schork (2000). "Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data." Am J Hum Genet 67(4): 947-59.
Falush, D., M. Stephens, et al. (2003). "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies." Genetics 164(4): 1567-87.
Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics 155(3): 1405-13.
Fearnhead, P. and N. G. Smith (2005). "A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes." Am J Hum Genet 77(5): 781-94.
Flint, J., R. M. Harding, et al. (1993). "The population genetics of the haemoglobinopathies." Baillieres Clin Haematol 6(1): 215-62.
Flint, J., R. M. Harding, et al. (1998). "The population genetics of the haemoglobinopathies." Baillieres Clin Haematol 11(1): 1-51.
Fortin, A., M. M. Stevenson, et al. (2002). "Susceptibility to malaria as a complex trait: big pressure from a tiny creature." Hum Mol Genet 11(20): 2469-78.
Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature 449(7164): 851-61.
Frodsham, A. J. and A. V. Hill (2004). "Genetics of infectious diseases." Hum Mol Genet 13 Spec No 2: R187-94.
Fu, Y. X. (1997). "Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection." Genetics 147(2): 915-25.

Fu, Y. X. and W. H. Li (1993). "Statistical tests of neutrality of mutations." Genetics **133**(3): 693-709.
Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-9.
Gambaro, G., F. Anglani, et al. (2000). "Association studies of genetic polymorphisms and complex disease." Lancet **355**(9200): 308-11.
Garcia, A., M. Cot, et al. (1998). "Genetic control of blood infection levels in human malaria: evidence for a complex genetic model." Am J Trop Med Hyg **58**(4): 480-8.
Garcia, A., S. Marquet, et al. (1998). "Linkage analysis of blood Plasmodium falciparum levels: interest of the 5q31-q33 chromosome region." Am J Trop Med Hyg **58**(6): 705-9.
Garrigan, D. and M. F. Hammer (2006). "Reconstructing human origins in the genomic era." Nat Rev Genet **7**(9): 669-80.
Garte, S. (2003). "Locus-specific genetic diversity between human populations: an analysis of the literature." Am J Human Biol **15**(6): 814-23.
Gendrel, D., M. Kombila, et al. (1991). "Protection against Plasmodium falciparum infection in children with hemoglobin S." Pediatr Infect Dis J **10**(8): 620-1.
Gillespie, J. H. (2000). "The neutral theory in an infinite population." Gene **261**(1): 11-8.
Glinka, S., L. Ometto, et al. (2003). "Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach." Genetics **165**(3): 1269-78.
Goldstein, D. B. and L. Chikhi (2002). "Human migrations and population structure: what we know and why it matters." Annu Rev Genomics Hum Genet **3**: 129-52.
Gonzalez-Neira, A., X. Ke, et al. (2006). "The portability of tagSNPs across populations: a worldwide survey." Genome Res **16**(3): 323-30.
Good, P. I. (2006). Resampling methods. A practical guide to data analysis, Birkhauser, Boston.
Greenwood, B. and T. Mutabingwa (2002). "Malaria in 2002." Nature **415**(6872): 670-2.
Greenwood, B. M. and J. R. Armstrong (1991). "Comparison of two simple methods for determining malaria parasite density." Trans R Soc Trop Med Hyg **85**(2): 186-8.
Greenwood, B. M., K. Bojang, et al. (2005). "Malaria." Lancet **365**(9469): 1487-98.
Greenwood, B. M., F. Groenendaal, et al. (1987). "Ethnic differences in the prevalence of splenomegaly and malaria in The Gambia." Ann Trop Med Parasitol **81**(4): 345-54.
Griffin, T. J. and L. M. Smith (2000). "Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry." Trends Biotechnol **18**(2): 77-84.
Gu, S., A. J. Pakstis, et al. (2007). "Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations." Eur J Hum Genet **15**(3): 302-12.
Guerra, C. A., P. W. Gikandi, et al. (2008). "The limits and intensity of Plasmodium falciparum transmission: implications for malaria control and elimination worldwide." PLoS Med **5**(2): e38.
Haldane, J. (1949). "The rate of mutation of human genes." Hereditas **Suppl 35**: 267-273.
Hamblin, M. T. and A. Di Rienzo (2000). "Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus." Am J Hum Genet **66**(5): 1669-79.
Hanchard, N. A., K. A. Rockett, et al. (2006). "Screening for recently selected alleles by analysis of human haplotype similarity." Am J Hum Genet **78**(1): 153-9.
HapMap (2004). "Integrating ethics and science in the International HapMap Project." Nat Rev Genet **5**(6): 467-75.
HapMap (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-320.

Hill, A. V. (1998). "The immunogenetics of human infectious diseases." Annu Rev Immunol **16**: 593-617.
Hill, A. V. (1999). "The immunogenetics of resistance to malaria." Proc Assoc Am Physicians **111**(4): 272-7.
Hill, A. V., C. E. Allsopp, et al. (1991). "Common west African HLA antigens are associated with protection from severe malaria." Nature **352**(6336): 595-600.
Hill, A. V., S. Bennett, et al. (1992). "HLA, malaria and dominant protective associations." Parasitol Today **8**(2): 57.
Hill, A. V., A. Jepson, et al. (1997). "Genetic analysis of host-parasite coevolution in human malaria." Philos Trans R Soc Lond B Biol Sci **352**(1359): 1317-25.
Hinds, D. A., L. L. Stuve, et al. (2005). "Whole-genome patterns of common DNA variation in three human populations." Science **307**(5712): 1072-9.
Hollox, E. J., M. Poulter, et al. (2001). "Lactase haplotype diversity in the Old World." Am J Hum Genet **68**(1): 160-172.
Hoogendoorn, B., S. L. Coleman, et al. (2003). "Functional analysis of human promoter polymorphisms." Hum Mol Genet **12**(18): 2249-54.
Hudson, R. R., K. Bailey, et al. (1994). "Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster." Genetics **136**(4): 1329-40.
Hudson, R. R., M. Kreitman, et al. (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics **116**(1): 153-9.
Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.
Hull, J., H. Ackerman, et al. (2001). "Unusual haplotypic structure of IL8, a susceptibility locus for a common respiratory virus." Am J Hum Genet **69**(2): 413-9.
Ioannidis, J. P., E. E. Ntzani, et al. (2001). "Replication validity of genetic association studies." Nat Genet **29**(3): 306-9.
Jeffreys, A. J., L. Kauppi, et al. (2001). "Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex." Nat Genet **29**(2): 217-22.
Jensen, J. D., Y. Kim, et al. (2005). "Distinguishing between selective sweeps and demography using DNA polymorphism data." Genetics **170**(3): 1401-10.
Jepson, A., F. Sisay-Joof, et al. (1997). "Genetic linkage of mild malaria to the major histocompatibility complex in Gambian children: study of affected sibling pairs." Bmj **315**(7100): 96-7.
Jepson, A. P., W. A. Banya, et al. (1995). "Genetic regulation of fever in Plasmodium falciparum malaria in Gambian twin children." J Infect Dis **172**(1): 316-9.
Johansson, A. and U. Gyllensten (2008). "Identification of local selective sweeps in human populations since the exodus from Africa." Hereditas **145**(3): 126-37.
Jorde, L. B. and S. P. Wooding (2004). "Genetic variation, classification and 'race'." Nat Genet **36**(11 Suppl): S28-33.
Joy, D. A., X. Feng, et al. (2003). "Early origin and recent expansion of Plasmodium falciparum." Science **300**(5617): 318-21.
Kaeuffer, R., D. Reale, et al. (2007). "Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium." Heredity **99**(4): 374-80.
Kaplan, N. L., R. R. Hudson, et al. (1989). "The "hitchhiking effect" revisited." Genetics **123**(4): 887-99.
Kayser, M., S. Brauer, et al. (2003). "A genome scan to detect candidate regions influenced by local natural selection in human populations." Mol Biol Evol **20**(6): 893-900.
Ke, X. and L. R. Cardon (2003). "Efficient selective screening of haplotype tag SNPs." Bioinformatics **19**(2): 287-8.

Ke, X., S. Hunt, et al. (2004). "The impact of SNP density on fine-scale patterns of linkage disequilibrium." Hum Mol Genet.
Khalil, E. A., E. E. Zijlstra, et al. (2002). "Epidemiology and clinical manifestations of Leishmania donovani infection in two villages in an endemic area in eastern Sudan." Trop Med Int Health **7**(1): 35-44.
Kidd, K. K., A. J. Pakstis, et al. (2004). "Understanding human DNA sequence variation." J Hered **95**(5): 406-20.
Kim, Y. and W. Stephan (2000). "Joint effects of genetic hitchhiking and background selection on neutral variation." Genetics **155**(3): 1415-27.
Kim, Y. and W. Stephan (2002). "Detecting a local signature of genetic hitchhiking along a recombining chromosome." Genetics **160**(2): 765-77.
Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(129): 624-6.
Kirchgatter, K. and H. A. Del Portillo (2005). "Clinical and molecular aspects of severe malaria." An Acad Bras Cienc **77**(3): 455-75.
Kosman, E. and K. J. Leonard (2007). "Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction." New Phytol **174**(3): 683-96.
Krishna, S., D. W. Waller, et al. (1994). "Lactic acidosis and hypoglycaemia in children with severe malaria: pathophysiological and prognostic significance." Trans R Soc Trop Med Hyg **88**(1): 67-73.
Kullo, I. J. and K. Ding (2007). "Patterns of population differentiation of candidate genes for cardiovascular disease." BMC Genet **8**: 48.
Kwiatkowski, D. (2000). "Genetic susceptibility to malaria getting complex." Curr Opin Genet Dev **10**(3): 320-4.
Kwiatkowski, D. and C. Bate (1995). "Inhibition of tumour necrosis factor (TNF) production by antimalarial drugs used in cerebral malaria." Trans R Soc Trop Med Hyg **89**(2): 215-6.
Kwiatkowski, D. P. (2005). "How malaria has affected the human genome and what human genetics can teach us about malaria." Am J Hum Genet **77**(2): 171-92.
Lewontin, R. C. (1964). "The interaction of selection and linkage. I. General considerations; heterotic models." Genetics **49**: 49-67.
Lewontin, R. C. and J. Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." Genetics **74**(1): 175-95.
Li, W. H. and L. A. Sadler (1991). "Low nucleotide diversity in man." Genetics **129**(2): 513-23.
Liu, H. X., L. Cartegni, et al. (2001). "A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes." Nat Genet **27**(1): 55-8.
Liu, N., S. L. Sawyer, et al. (2004). "Haplotype block structures show significant variation among populations." Genet Epidemiol **27**(4): 385-400.
Luoni, G., J. Forton, et al. (2005). "Population-specific patterns of linkage disequilibrium in the human 5q31 region." Genes Immun **6**(8): 723-7.
Luoni, G., F. Verra, et al. (2001). "Antimalarial antibody levels and IL4 polymorphism in the Fulani of West Africa." Genes Immun **2**(7): 411-4.
Mackinnon, M. J., T. W. Mwangi, et al. (2005). "Heritability of malaria in Africa." PLoS Med **2**(12): e340.
MalariaGEN (2008). "A global network for investigating the genomic epidemiology of malaria." Nature **456**(7223): 732-7.
Mangano, V. D., G. Luoni, et al. (2008). "Interferon regulatory factor-1 polymorphisms are associated with the control of Plasmodium falciparum infection." Genes Immun **9**(2): 122-9.

Manica, A., F. Prugnolle, et al. (2005). "Geography is a better determinant of human genetic differentiation than ethnicity." Hum Genet 118(3-4): 366-71.
Marchini, J., D. Cutler, et al. (2006). "A comparison of phasing algorithms for trios and unrelated individuals." Am J Hum Genet 78(3): 437-50.
Marchini, J., B. Howie, et al. (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." Nat Genet 39(7): 906-13.
Marquet, S., L. Abel, et al. (1996). "Genetic localization of a locus controlling the intensity of infection by Schistosoma mansoni on chromosome 5q31-q33." Nat Genet 14(2): 181-4.
Marsh, D. G., J. D. Neely, et al. (1994). "Linkage analysis of IL4 and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations." Science 264(5162): 1152-6.
Marsh, K., M. English, et al. (1996). "The pathogenesis of severe malaria in African children." Ann Trop Med Parasitol 90(4): 395-402.
Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proc Natl Acad Sci U S A 74(2): 560-4.
McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature 351(6328): 652-4.
McGuire, W., A. V. Hill, et al. (1994). "Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria." Nature 371(6497): 508-10.
McVean, G., C. C. Spencer, et al. (2005). "Perspectives on human genetic variation from the HapMap Project." PLoS Genet 1(4): e54.
McVean, G. A., S. R. Myers, et al. (2004). "The fine-scale structure of recombination rate variation in the human genome." Science 304(5670): 581-4.
Meyers, D. A., D. S. Postma, et al. (1994). "Evidence for a locus regulating total serum IgE levels mapping to chromosome 5." Genomics 23(2): 464-70.
Michelson, A. D., J. D. Wencel-Drake, et al. (1994). "Platelet activation results in a redistribution of glycoprotein IV (CD36)." Arterioscler Thromb 14(7): 1193-201.
Miller, E. N., M. Fadl, et al. (2007). "Y chromosome lineage- and village-specific genes on chromosomes 1p22 and 6q27 control visceral leishmaniasis in Sudan." PLoS Genet 3(5): e71.
Miller, L. H., D. I. Baruch, et al. (2002). "The pathogenic basis of malaria." Nature 415(6872): 673-9.
Miller, L. H., M. F. Good, et al. (1994). "Malaria pathogenesis." Science 264(5167): 1878-83.
Modiano, D., A. Chiucchiuini, et al. (1998). "Humoral response to Plasmodium falciparum Pf155/ring-infected erythrocyte surface antigen and Pf332 in three sympatric ethnic groups of Burkina Faso." Am J Trop Med Hyg 58(2): 220-4.
Modiano, D., G. Luoni, et al. (2001). "The lower susceptibility to Plasmodium falciparum malaria of Fulani of Burkina Faso (west Africa) is associated with low frequencies of classic malaria-resistance genes." Trans R Soc Trop Med Hyg 95(2): 149-52.
Modiano, D., G. Luoni, et al. (2001). "Haemoglobin C protects against clinical Plasmodium falciparum malaria." Nature 414(6861): 305-8.
Modiano, D., V. Petrarca, et al. (1996). "Different response to Plasmodium falciparum malaria in west African sympatric ethnic groups." Proc Natl Acad Sci U S A 93(23): 13206-11.
Molyneux, M. E., T. E. Taylor, et al. (1989). "Clinical features and prognostic indicators in paediatric cerebral malaria: a study of 131 comatose Malawian children." Q J Med 71(265): 441-59.

Mu, J., P. Awadalla, et al. (2005). "Recombination hotspots and population structure in Plasmodium falciparum." PLoS Biol **3**(10): e335.
Nagel, R. L. and A. F. Fleming (1992). "Genetic epidemiology of the beta s gene." Baillieres Clin Haematol **5**(2): 331-65.
Neel, J. V. (1962). "Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?" Am J Hum Genet **14**: 353-62.
Nei, M. (1973). "Analysis of gene diversity in subdivided populations." Proc Natl Acad Sci U S A **70**(12): 3321-3.
Nei, M. (1978). "Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals." Genetics **89**(3): 583-590.
Nei, M. (1982). "Evolution of human races at the gene level." Prog Clin Biol Res **103 Pt A**: 167-81.
Nei, M. and M. W. Feldman (1972). "Identity of genes by descent within and between populations under mutation and migration pressures." Theor Popul Biol **3**(4): 460-5.
Nielsen, R. (2001). "Statistical tests of selective neutrality in the age of genomics." Heredity **86**(Pt 6): 641-7.
Niu, T. (2004). "Algorithms for inferring haplotypes." Genet Epidemiol **27**(4): 334-47.
Nobrega, M. A., I. Ovcharenko, et al. (2003). "Scanning human gene deserts for long-range enhancers." Science **302**(5644): 413.
Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-6.
Ohashi, J., I. Naka, et al. (2004). "Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection." Am J Hum Genet **74**(6): 1198-208.
Organisation, W. H. (2000). "Severe falciparum malaria. World Health Organization, Communicable Diseases Cluster." Trans R Soc Trop Med Hyg **94 Suppl 1**: S1-90.
Orjih, A. U., R. Chevli, et al. (1985). "Toxic heme in sickle cells: an explanation for death of malaria parasites." Am J Trop Med Hyg **34**(2): 223-7.
Paabo, S. (2003). "The mosaic that is our genome." Nature **421**(6921): 409-12.
Pagnier, J., J. G. Mears, et al. (1984). "Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa." Proc Natl Acad Sci U S A **81**(6): 1771-3.
Pasvol, G. (2006). "The treatment of complicated and severe malaria." Br Med Bull **75-76**: 29-47.
Patil, N., A. J. Berno, et al. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." Science **294**(5547): 1719-23.
Payseur, B. A., A. D. Cutter, et al. (2002). "Searching for evidence of positive selection in the human genome using patterns of microsatellite variability." Mol Biol Evol **19**(7): 1143-53.
Phillips, M. S., R. Lawrence, et al. (2003). "Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots." Nat Genet **33**(3): 382-7.
Piazza, A., W. R. Mayr, et al. (1985). "Genetic and population structure of four Sardinian villages." Ann Hum Genet **49 (Pt 1)**: 47-63.
Plagnol, V. and J. D. Wall (2006). "Possible ancestral structure in human populations." PLoS Genet **2**(7): e105.
Postma, D. S., E. R. Bleecker, et al. (1995). "Genetic susceptibility to asthma--bronchial hyperresponsiveness coinherited with a major gene for atopy." N Engl J Med **333**(14): 894-900.
Powars, D. and A. Hiti (1993). "Sickle cell anemia. Beta s gene cluster haplotypes as genetic markers for severe disease expression." Am J Dis Child **147**(11): 1197-202.
Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: models and data." Am J Hum Genet **69**(1): 1-14.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-59.
Pritchard, J. K., M. Stephens, et al. (2000). "Association mapping in structured populations." Am J Hum Genet **67**(1): 170-81.
Przeworski, M. (2002). "The signature of positive selection at randomly chosen loci." Genetics **160**(3): 1179-89.
Przeworski, M., R. R. Hudson, et al. (2000). "Adjusting the focus on human variation." Trends Genet **16**(7): 296-302.
Ptak, S. E., D. A. Hinds, et al. (2005). "Fine-scale recombination patterns differ between chimpanzees and humans." Nat Genet **37**(4): 429-34.
Rahimi, Z., M. Karimi, et al. (2003). "Beta-globin gene cluster haplotypes in sickle cell patients from southwest Iran." Am J Hematol **74**(3): 156-60.
Ramirez-Soriano, A., O. Lao, et al. (2005). "Haplotype tagging efficiency in worldwide populations in CTLA4 gene." Genes Immun **6**(8): 646-57.
Reich, D. E., M. Cargill, et al. (2001). "Linkage disequilibrium in the human genome." Nature **411**(6834): 199-204.
Reyburn, H., R. Mbatia, et al. (2005). "Association of transmission intensity and age with clinical manifestations and case fatality of severe Plasmodium falciparum malaria." Jama **293**(12): 1461-70.
Richie, T. L. and A. Saul (2002). "Progress and challenges for malaria vaccines." Nature **415**(6872): 694-701.
Rihet, P., L. Abel, et al. (1998). "Human malaria: segregation analysis of blood infection levels in a suburban area and a rural area in Burkina Faso." Genet Epidemiol **15**(5): 435-50.
Rihet, P., Y. Traore, et al. (1998). "Malaria in humans: Plasmodium falciparum blood infection levels are linked to chromosome 5q31-q33." Am J Hum Genet **63**(2): 498-505.
Riley, E. M., O. Olerup, et al. (1992). "MHC and malaria: the relationship between HLA class II alleles and immune responses to Plasmodium falciparum." Int Immunol **4**(9): 1055-63.
Rioux, J. D., M. J. Daly, et al. (2001). "Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease." Nat Genet **29**(2): 223-8.
Roberts, D. J. and T. N. Williams (2003). "Haemoglobinopathies and resistance to malaria." Redox Rep **8**(5): 304-10.
Rogers, J. (1972). "Measures of genetic similarity and genetic distance." Studies in genetics(Austin, TX, USA: University of Texas): 145–143.
Rosenberg, N. A., T. Burke, et al. (2001). "Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds." Genetics **159**(2): 699-713.
Rosenberg, N. A., J. K. Pritchard, et al. (2002). "Genetic structure of human populations." Science **298**(5602): 2381-5.
Roth, E. F., Jr., M. Friedman, et al. (1978). "Sickling rates of human AS red cells infected in vitro with Plasmodium falciparum malaria." Science **202**(4368): 650-2.
Sabeti, P. C., D. E. Reich, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-7.
Sabeti, P. C., S. F. Schaffner, et al. (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-20.
Sabeti, P. C., P. Varilly, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-8.

Sachs, J. and P. Malaney (2002). "The economic and social burden of malaria." Nature **415**(6872): 680-5.
Salisbury, B. A., M. Pungliya, et al. (2003). "SNP and haplotype variation in the human genome." Mutat Res **526**(1-2): 53-61.
Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-7.
Saunders, M. A., M. F. Hammer, et al. (2002). "Nucleotide variability at G6pd and the signature of malarial selection in humans." Genetics **162**(4): 1849-61.
Saunders, M. A., M. Slatkin, et al. (2005). "The extent of linkage disequilibrium caused by selection on G6PD in humans." Genetics **171**(3): 1219-29.
Sawyer, S. L., N. Mukherjee, et al. (2005). "Linkage disequilibrium patterns vary substantially among populations." Eur J Hum Genet **13**(5): 677-86.
Sawyer, S. L., L. I. Wu, et al. (2005). "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." Proc Natl Acad Sci U S A **102**(8): 2832-7.
Saxena, R., B. F. Voight, et al. (2007). "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." Science **316**(5829): 1331-6.
Schork, N. J. (2002). "Power calculations for genetic association studies using estimated probability distributions." Am J Hum Genet **70**(6): 1480-9.
Service, S., J. DeYoung, et al. (2006). "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies." Nat Genet **38**(5): 556-60.
Shifman, S. and A. Darvasi (2001). "The value of isolated populations." Nat Genet **28**(4): 309-10.
Simonsen, K. L., G. A. Churchill, et al. (1995). "Properties of statistical tests of neutrality for DNA polymorphism data." Genetics **141**(1): 413-29.
Singh, B., L. Kim Sung, et al. (2004). "A large focus of naturally acquired Plasmodium knowlesi infections in human beings." Lancet **363**(9414): 1017-24.
Sjoberg, K., J. P. Lepers, et al. (1992). "Genetic regulation of human anti-malarial antibodies in twins." Proc Natl Acad Sci U S A **89**(6): 2101-4.
Slatkin, M. (1994). "Linkage disequilibrium in growing and stable populations." Genetics **137**(1): 331-6.
Smith, T. A., R. Leuenberger, et al. (2001). "Child mortality and malaria transmission intensity in Africa." Trends Parasitol **17**(3): 145-9.
Snow, R. W., C. A. Guerra, et al. (2005). "The global distribution of clinical episodes of Plasmodium falciparum malaria." Nature **434**(7030): 214-7.
Sorensen, T. I., G. G. Nielsen, et al. (1988). "Genetic and environmental influences on premature death in adult adoptees." N Engl J Med **318**(12): 727-32.
Steinberg, M. H., Z. H. Lu, et al. (1998). "Hematological effects of atypical and Cameroon beta-globin gene haplotypes in adult sickle cell anemia." Am J Hematol **59**(2): 121-6.
Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." Am J Hum Genet **73**(5): 1162-9.
Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-89.
Stevenson, M. M., J. J. Lyanga, et al. (1982). "Murine malaria: genetic control of resistance to Plasmodium chabaudi." Infect Immun **38**(1): 80-8.
Stevenson, M. M. and E. Skamene (1985). "Murine malaria: resistance of AXB/BXA recombinant inbred mice to Plasmodium chabaudi." Infect Immun **47**(2): 452-6.

Stirnadel, H. A., H. P. Beck, et al. (1999). "Heritability and segregation analysis of immune responses to specific malaria antigens in Papua New Guinea." Genet Epidemiol **17**(1): 16-34.

Storz, J. F., B. A. Payseur, et al. (2004). "Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa." Mol Biol Evol **21**(9): 1800-11.

Su, X. Z., J. Mu, et al. (2003). "The "Malaria's Eve" hypothesis and the debate concerning the origin of the human malaria parasite Plasmodium falciparum." Microbes Infect **5**(10): 891-6.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-95.

Tarazona-Santos, E. and S. A. Tishkoff (2005). "Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus." Genes Immun **6**(1): 53-65.

Teo, Y. Y., M. Inouye, et al. (2008). "Whole genome-amplified DNA: insights and imputation." Nat Methods **5**(4): 279-80.

Teo, Y. Y., M. Inouye, et al. (2007). "A genotype calling algorithm for the Illumina BeadArray platform." Bioinformatics **23**(20): 2741-6.

Terrenato, L., S. Shrestha, et al. (1988). "Decreased malaria morbidity in the Tharu people compared to sympatric populations in Nepal." Ann Trop Med Parasitol **82**(1): 1-11.

Terwilliger, J. D. and K. M. Weiss (1998). "Linkage disequilibrium mapping of complex disease: fantasy or reality?" Curr Opin Biotechnol **9**(6): 578-94.

Tiret, L., O. Poirier, et al. (2002). "Heterogeneity of linkage disequilibrium in human genes has implications for association studies of common diseases." Hum Mol Genet **11**(4): 419-29.

Tishkoff, S. A., E. Dietzsch, et al. (1996). "Global patterns of linkage disequilibrium at the CD4 locus and modern human origins." Science **271**(5254): 1380-7.

Tishkoff, S. A., R. Varkonyi, et al. (2001). "Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance." Science **293**(5529): 455-62.

Tishkoff, S. A. and B. C. Verrelli (2003). "Patterns of human genetic diversity: implications for human evolutionary history and disease." Annu Rev Genomics Hum Genet **4**: 293-340.

Tongren, J. E., F. Zavala, et al. (2004). "Malaria vaccines: if at first you don't succeed." Trends Parasitol **20**(12): 604-10.

Toomajian, C., R. S. Ajioka, et al. (2003). "A method for detecting recent selection in the human genome from allele age estimates." Genetics **165**(1): 287-97.

Vallender, E. J. and B. T. Lahn (2004). "Positive selection on the human genome." Hum Mol Genet **13 Spec No 2**: R245-54.

Vivenes De Lugo, M., A. Rodriguez-Larralde, et al. (2003). "Beta-globin gene cluster haplotypes as evidence of African gene flow to the northeastern coast of Venezuela." Am J Hum Biol **15**(1): 29-37.

Voight, B. F., S. Kudaravalli, et al. (2006). "A map of recent positive selection in the human genome." PLoS Biol **4**(3): e72.

Voight, B. F. and J. K. Pritchard (2005). "Confounding from cryptic relatedness in case-control association studies." PLoS Genet **1**(3): e32.

Watkins, W. S., A. R. Rogers, et al. (2003). "Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms." Genome Res **13**(7): 1607-18.

Weatherall, D. J., L. H. Miller, et al. (2002). "Malaria and the red cell." Hematology (Am Soc Hematol Educ Program): 35-57.

Weiss, K. M. and J. D. Terwilliger (2000). "How many diseases does it take to map a gene with SNPs?" Nat Genet **26**(2): 151-7.

White, N. J. (1987). "Clinical and pathological aspects of severe malaria." Acta Leiden **56**: 27-46.

Wilkinson, R. J. and G. Pasvol (1997). "Host resistance to malaria runs into swampy water." Trends Microbiol **5**(6): 213-5.

Williams, T. N., T. W. Mwangi, et al. (2005). "Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases." J Infect Dis **192**(1): 178-86.

Williamson, S. H., R. Hernandez, et al. (2005). "Simultaneous inference of selection and population growth from patterns of variation in the human genome." Proc Natl Acad Sci U S A **102**(22): 7882-7.

Winstanley, P. A., S. A. Ward, et al. (2002). "Clinical status and implications of antimalarial drug resistance." Microbes Infect **4**(2): 157-64.

Wood, E. T., D. A. Stover, et al. (2005). "The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria." Am J Hum Genet **77**(4): 637-42.

Yu, F., P. C. Sabeti, et al. (2005). "Positive selection of a pre-expansion CAG repeat of the human SCA2 gene." PLoS Genet **1**(3): e41.

Zago, M. A., W. A. Silva, Jr., et al. (2000). "Atypical beta(s) haplotypes are generated by diverse genetic mechanisms." Am J Hematol **63**(2): 79-84.

Zago, M. A., W. A. Silva, Jr., et al. (2001). "Rearrangements of the beta-globin gene cluster in apparently typical betaS haplotypes." Haematologica **86**(2): 142-5.

Zaykin, D. V., Z. Meng, et al. (2006). "Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method." Am J Hum Genet **78**(5): 737-46.

Zeggini, E., W. Rayner, et al. (2005). "An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets." Nat Genet **37**(12): 1320-2.

Zhang, K., P. Calabrese, et al. (2002). "Haplotype block structure and its applications to association studies: power and study designs." Am J Hum Genet **71**(6): 1386-94.

Zhang, L., X. Cui, et al. (1992). "Whole genome amplification from a single cell: implications for genetic analysis." Proc Natl Acad Sci U S A **89**(13): 5847-51.

Zondervan, K. T. and L. R. Cardon (2004). "The complex interplay among factors that influence allelic association." Nat Rev Genet **5**(2): 89-100.

Zuckerkandl, E. (2002). "Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine." Genetica **115**(1): 105-29.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Rogers JS. 1972. Measures of genetic similarity and genetic distance, pp. 145–143. In: *Studies in genetics*. Austin, TX, USA: University of Texas.

Schnieder, S., Roessli, D. and Excoffier, L. (2000) Arlequin ver.2.000:A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland. http://anthro.unige.ch/arlequin.

Appendix 1: Ethical approval and informed consent.

Figure1: Ethical Approval from the Ethical Committee of the Institute of Endemic Diseases for the collection of the Sudanese samples.

بسم الله الرحمن الرحيم

INSTITUTE OF ENDEMIC DISEASES
UNIVERSITY OF KHARTOUM
MEDICAL CA1MPUS
P.O BOX 102, KHARTOUM, SUDAN
TEL: +249-11-779712/ 793262
FAX: +249-11-779712/793263

معهد الأمراض المتوطنة
جامعة الخرطوم
مجمع العلوم الطبية
ص.ب ١٠٢ ، خرطوم-سودان
ت:١١-٧٧٩٧١٢-٢٤٩+—٧٩٣٢٦٢
فاكس:١١-٧٧٩٧١٢-٢٤٩+—٧٩٣٢٦٢

Institutional Endorsement

This is to confirm that the Research proposal/Project entitled

The Genetic resistance to malaria in two communities of different ethnic origin in eastern Sudan

Has been approved by the scientific committee of the Institute at its meeting dated 15 / 1 / 2004

Muntaser E. Ibrahim
(Secretary)

Signature .....

Date 22-1-2004

Moawia M. Mukhtar
(Chairman)

Signature .....

Figure 2: Informed consent form for the Sudanese samples.

اقرار بالموافقة

ياجماعة نحن اطباء و باحثين من جامعة الخرطوم معهد الأمراض المتوطنة. جيناكم دايرين نعرف السبب ليه الملاريا بتجي لي ناس و ناس لا و كيف تاريخ المجموعة البشرية ممكن يؤثر على انتشار الملاريا و دور العامل الوراثى وعشان و أطباء وهم د. منتصر الطيب ابراهيم كدة عايزين نعمل دراسة فى الموضوع دا. المسؤلون من هذة الدراسة هم باحثين و د. دومنيك كويتنكوسكي من جامعة اكسفورد .
حناخد عينة من الخشم بفرشة الأسنان وعينة دم من الاصبع و بعضكم حناخد منه عينة دم من الوريد اذا في أي مرض حنعالجكم منه وكمان العلاج حيكون لكل زول في المنطقة حتى لو الزول مادايز يشارك معانا في الدراسة دي0 وأهم ولو تم اكتشاف ,حاجة ياجماعة عايزنكم تعرفوها هي انو نتيجة الفحص بتاعكم مافي اي زول حبعرفها الا سيد العينة علمي من العينات دي ما حيكون عندو اي قيمة تجارية بدون موافقتكم0 الزول كان موافق يقوم يكتب لينا اسمه ويمضي او يبصم

Consent

We are a group of doctors and researchers from the Institute of Endemic Diseases, University of Khartoum. Our interest is on the question of why malaria affects some human populations more than others and the relationship of the population genetic history and population structure on the propagation of diseases.

The persons in charge of this study are Dr. Muntaser E. Ibrahim from the University of Khartoum and Dr. Dominic Kwiatkowski from Oxford University (UK). We will be taking samples of buccal cells for DNA analysis and we will take blood from a finger prick. From some of you we will be taking 2cc of venous blood. We will be looking at your health complaints and provide medical treatment for all in need even those who are not willing to participate in the study. Most importantly, all your personal information will be kept confidential and any scientific outcome from the study will not be commercialized without your approval; and consent.

I agree/ my child to participate أوافق علي مشاركتي/ مشاركة طفلي

Name الأسم
0000000000000000000000000000000000

Signature الأمضاء

0000000000000000000000000000

العناوين----------

Appendix 2: Haplotypes sequences in the 5q31 and HBB regions.

Table 1: Haplotypes of the 5q31 region in the Hausa sample.

Haplotype ID	Haplotype sequence	Haplotype Frequency
1	12111121111111111111221111111	1.000000
2	12111121111111211122111111211	1.000000
3	12111122111111111122112112211	2.000000
4	12111112112111111111111111111	2.000000
5	12111112111111111122112112211	1.000000
6	12112222111111111111121111211	1.000000
7	12112211112111111111111111211	1.000000
8	12112212222111111111112112112	1.000000
9	12112212222111111122111111211	1.000000
10	12112212222111111122112121212	2.000000
11	12112212222111111122112121212	1.000000
12	12112212112111111111111111122	1.000000
13	12112212112111111111111111211	1.000000
14	12112212112111111111121111211	1.000000
15	12112212111111111111211112112	1.000000
16	12112212111111111111211112122	2.000000
17	12112212111111111111221111111	4.000000
18	12112212111111111122122111111	1.000000
19	12112212111111111212121111111	1.000000
20	12212121111111111111112112211	1.000000
21	12212122222111111111221111111	1.000000
22	12212122111111111111122112211	1.000000
23	12212122111111111111211111111	1.000000
24	12212122111111111111211112112	1.000000
25	12212122111111111111211112122	3.000000
26	12212122111111111111212111111	1.000000
27	12212122111111111111221111111	2.000000
28	12212122111111112111121111211	1.000000
29	12212111222111211122122122122	1.000000
30	12212111222111211122122122211	1.000000
31	12212111112111111111111111211	1.000000
32	12212111111111112111221111111	2.000000
33	12212112222111112111111111112	1.000000
34	12212112222111112111111112111	1.000000
35	12221121222111111111111112112	2.000000
36	12221121222111111122112112211	1.000000
37	12221122112111111111111112111	4.000000
38	12221112111111111122112112211	1.000000
39	11111121111111111111111111211	1.000000
40	11111121111111111111111112112	1.000000
41	11111121111111112111121111111	9.000000
42	11111122111111111111111211122	1.000000
43	11111221222111211122111111111	1.000000
44	11111221221111111122112112122	2.000000

45	1111122111211121112211121122211	2.000000
46	1111122222111111111212121111111	1.000000
47	1111222222211111112211212112121 2	1.000000
48	1111221111211111111111111111211	1.000000
49	1111221111111111111111111112211	2.000000
50	1111221111111212222211112122 11	1.000000
51	1112122222211111112211212112121 2	1.000000
52	1112121211211121112211111111211	1.000000
53	1121211111211111111111111111211	1.000000
54	1122112111111111111111111111112	1.000000
55	1122112222111111111221221111111	1.000000
56	1122112222111111211111112112211	1.000000
57	1122112211211111111111111112111	2.000000
58	2221211112211111111212221211211	1.000000
59	2111112111111111211112211111111	1.000000
60	2111122122211111111111111211212	1.000000
61	2111121211221212222211111112112	1.000000
62	2111212111111111111221112112112	1.000000
63	2111212111111111111221112112211	1.000000
64	2111211122211111212212111111211	2.000000
65	2111222122111111111221112111211	1.000000
66	2111222122111111111221112112112	1.000000
67	2111222122111111111211221111111	2.000000
68	2111222211111111111111211111111	3.000000
69	2111222211111111111111211111211	1.000000
70	2111222211111111111221222112211	1.000000
71	2111221111211111111111111111211	2.000000
72	2111221222211111111111121222211	1.000000
73	2111221222211111211111221111111	1.000000
74	2111221211211111111111112111211	2.000000
75	2111221211211111111221112111211	1.000000
76	2111221211221212222211111112112	2.000000
77	2111221211221212222211111112122	1.000000
78	2111221211221212222211112112111	1.000000
79	2111221211221212222211112112212	1.000000
80	2111221211111111111221121221212	1.000000
81	2121112111111111111221221111211	1.000000
82	2121212122111111111221111112111	1.000000
83	2121212122111111111221112111211	1.000000
84	2121212111111111111221112112111	1.000000
85	2121212111111111111221112112112	3.000000
86	2121212111111111111221112112122	1.000000
87	2121212111111111111211212111112	1.000000
88	2121212111111111211111221111111	1.000000
89	2121212222211111111111111111112	1.000000
90	2121212222211111111111211111111	1.000000
91	2121212222111111111221111212112	1.000000
92	2121212211111111111221112112111	1.000000
93	2121212211111111111221112112112	1.000000
94	2121212211111111111221221111211	1.000000
95	2121212211111111111221221112211	1.000000
96	2121211122211121112212212222 11	2.000000

97	21212111112211111111111111211	1.000000
98	21212111111111111111221111111	1.000000
99	21212111111111111221222112211	1.000000
100	21212111111111111221222211211	1.000000
101	21212111111111121111221111111	1.000000
102	21212112111111111111221111111	1.000000
103	21212112111111111221222112211	1.000000
104	21212222112111112122112112211	1.000000
105	21212222111111111221121211211	1.000000
106	21221121111111111221112112111	1.000000
107	21221121111111112122112112211	2.000000
108	21221121111111122222122112211	1.000000
109	21221122222111111112122211211	1.000000
110	21221122221111112111112112211	1.000000
111	21221122112111211111221111111	1.000000
112	21221111222111211111111111211	1.000000
113	21221111222111211122111111111	5.000000
114	21221111222111211122111111112	1.000000
115	21221111222111211122122111111	1.000000
116	21221111111111111111112112211	1.000000
117	21221111111111111111112112212	1.000000
118	21221112222111111111111111111	2.000000
119	21221112222111111111112112111	1.000000
120	21221112222111112122122112211	1.000000
121	21221112222111122222122112211	1.000000
122	21221112112111211122111212211	1.000000
123	21221112111111112222112122212	1.000000
124	21221221221111111111111111211	1.000000
125	21221222222111111111111212111	1.000000
126	21221222222111111111211112122	1.000000
127	21221212111111112221122111111	1.000000

Table 2: Haplotypes of the 5q31 region in the Masalit sample.

Haplotype ID	Haplotype sequence	Haplotype Frequency
1	21112211111111111111111212211	2.000000
2	21112211111111111111211112122	1.000000
3	21112211111111111111212112211	1.000000
4	21112211111111111111221111111	1.000000
5	21112211112111112122121111112	1.000000
6	21112212111111111111221211211	1.000000
7	21112212112111111111111111211	1.000000
8	21112112112111111111121111211	1.000000
9	21111212111111111221212112211	1.000000
10	21111111112111111111111111211	1.000000
11	21212112111111111111121111211	2.000000
12	21212112111111111111221111111	2.000000
13	21212121111111111211111112211	1.000000
14	21212121111111111221112112211	1.000000
15	21221221112111111111111112211	1.000000
16	21221221112111111122111212211	1.000000
17	21221221112111211111111112111	1.000000
18	21221221221111111111221111111	1.000000
19	21221111111111111111211112211	1.000000
20	21221111212111211122111212112	1.000000
21	21221111222111211122111111211	2.000000
22	21221112111111211122111112211	1.000000
23	21221121112111211111121111111	1.000000
24	21221121112111211122111212211	8.000000
25	21221121122111211122111111211	1.000000
26	21221121222111111122112112211	3.000000
27	21221122111111111111221111111	1.000000
28	21221122221111111122112112211	2.000000
29	21221122222111111111111111112	1.000000
30	11112211112111111122111112211	1.000000
31	11112212111111111111111112122	2.000000
32	11112212111111111211211211111	1.000000
33	11112212111111111221111112211	2.000000
34	11112212111111111221111212211	1.000000
35	11112212111111111221112111112	1.000000
36	11112111112111212111121211212	1.000000
37	11111211111111111111221111111	2.000000
38	11111211112111112111111112111	2.000000
39	11111211112111211111211212211	1.000000
40	11111211222111211122112121111	1.000000
41	11111211222111211122121212211	1.000000
42	11111212111111111111111111211	2.000000
43	11111212111111111111111212211	2.000000
44	11111212111111111111221111111	11.000000
45	11111212111111111111221111112	1.000000
46	11111212111111111111221112211	1.000000
47	11111212111111111112111111111	1.000000

48	11111212111111111112211121122 11	1.000000
49	11111212111111111122121211221 1	1.000000
50	11111212111111112111122111111 1	1.000000
51	11111212111111211122111112111 1	1.000000
52	11111212112111111111111111111 1	1.000000
53	11111212112212122222111111211 1	1.000000
54	11111212112212122222111211221 1	3.000000
55	11111212112212122222111211211 2	1.000000
56	11111221111111112111122111111 1	1.000000
57	11111221111111112221121211221 1	1.000000
58	11111221111111112221122111221 1	1.000000
59	11111221222111111111111111111 2	2.000000
60	11111221222111112111111111111 2	1.000000
61	11111222112111211122111121111 1	1.000000
62	11111111111111111111112111111 1	1.000000
63	11111111222111111122111211221 1	1.000000
64	11111112111111111111122111111 1	2.000000
65	11111112111111111122122111211 2	1.000000
66	11111122111111112111122111121 2	1.000000
67	11111122111111112111122111111 1	1.000000
68	11111122111111112122111121211 1	1.000000
69	11121212112111111111111111211 1	1.000000
70	11121221111111112111111112111 1	1.000000
71	11212121111111111111112111112 11	1.000000
72	11212121111111111111122111111 1	3.000000
73	11212121111111111122111111111 1	1.000000
74	11212121111111111122111211221 1	1.000000
75	11212121111111111122121211221 1	1.000000
76	11212121222111211122112122221 1	1.000000
77	11212122111111112111122121121 1	1.000000
78	11211212111111111111122111111 1	1.000000
79	11221222111111111122112111111 1	1.000000
80	11221121111111112211112121111 1	1.000000
81	12112211221111111122111121221 1	1.000000
82	12112212111111111111111111111 1	1.000000
83	12112212111111111111112112121 1	1.000000
84	12112111121111111111111111211 1	1.000000
85	12112111121111111111111112111 1	1.000000
86	12112111121111111111111212211 1	1.000000
87	12112112111111111112121111111 1	2.000000
88	12112112111111111122111211112 1	1.000000
89	12112121112111112211112121111 1	1.000000
90	12111112111111111111111111211 1	1.000000
91	12111112111111111111112112211 1	1.000000
92	12111121111111111111112111112 1	1.000000
93	12111121111111111111122111111 1	1.000000
94	12111121111111111111122112211 1	1.000000
95	12111121111111111112111211211 1	1.000000
96	12111121222111112122112111111 1	1.000000
97	12111122222111111111112111111 1	1.000000
98	12212111111111111111112112211 1	1.000000
99	12212111111111111111112121122 11	1.000000

100	122121111111111111111221111111	2.000000
101	12212111111111111221111111111	1.000000
102	12212111111111111221112111211	1.000000
103	12212111111111111221221112112	1.000000
104	12212111111111122111112112211	1.000000
105	12212112111111111111221112112	1.000000
106	12212112111111112111221111111	1.000000
107	12212112112111112111111112111	1.000000
108	12212112112111112122121212112211	1.000000
109	12212121112111111111221111111	1.000000
110	12212122111111112111121111111	2.000000
111	12212122111111112111221111111	1.000000
112	12212122111111112111222112211	1.000000
113	12212122222111111111111111112	1.000000
114	12211121222111111112112211111111	1.000000
115	12221112111111111111221211112	1.000000
116	12221112111111111122121212112211	1.000000
117	12221122111111111111211111111	1.000000

Table 3: Haplotypes of the HBB region in the combined Hausa and Masalit samples.

Haplotype ID	Haplotype sequence	Haplotype Frequency
1	1111111212122211222111111 2	1.000000
2	1111111212121121111112111 2	1.000000
3	1111111212112111111111112	2.000000
4	1111111212112111111112112	1.000000
5	1111111212112111111212111 1	1.000000
6	1111111212112111212112111 1	1.000000
7	1111111212112111212212111 1	2.000000
8	1111111212112112212111112	1.000000
9	1111111212112112212111221 1	1.000000
10	1111111212112211222111112	1.000000
11	1111111212111121111211111	3.000000
12	1111111212111121111221111	1.000000
13	1111111212111121111221211	1.000000
14	1111111212111211222121111 2	1.000000
15	1111111212222111212111211	1.000000
16	1111111212212111111121211	1.000000
17	1111111212212112212111111	1.000000
18	1111111212212112212111212	1.000000
19	1111111212212211222111112	1.000000
20	1111121212112111111121211	2.000000
21	1111121212112111211121112	1.000000
22	1111121212112112212111211	1.000000
23	1111121212112112212112211	2.000000
24	1111121212112121111211211	1.000000
25	1111121212111121111212111	1.000000
26	1111121212111121111221111	2.000000
27	1111121212222111212111211	1.000000
28	1111121212212111111212111	1.000000
29	1111121212212112212111112	1.000000
30	1111121222212112212112112	1.000000
31	1111122212112112212112121 1	1.000000
32	1111211212122111111112211	1.000000
33	1111211212122111111212111	1.000000
34	1111211212122111111212111 2	1.000000
35	1111211212122111111212112	1.000000
36	1111211212122111111212112	2.000000
37	1111211212122211222111211	1.000000
38	1111211212121111111111111	2.000000
39	1111211212121121111221211	1.000000
40	1111211212111121111111211	1.000000
41	1111211212111211222121111	1.000000
42	1111211212222111111111211	1.000000
43	1111211212212111212111211	1.000000
44	1111211222122111111212111	2.000000
45	1111211222122111111212112	6.000000
46	1111211222122111111212211	3.000000
47	1111212212122111111212112	2.000000

48	111121221212221122212121111	1.000000
49	111121221212221122212121211	1.000000
50	111121221212112111112211111	1.000000
51	111121221211211111111112111	1.000000
52	111121221211211111111112211	1.000000
53	111121221211211221211121111	1.000000
54	111121221211212111111221211	1.000000
55	111121221211112111111111211	1.000000
56	111121221211112111111211111	1.000000
57	111121221211112111111211112	1.000000
58	111121221211112111111211211	1.000000
59	111121221211112111111221111	1.000000
60	111121221222212111111111221	1.000000
61	111121221222221122212211111	1.000000
62	111121221221211111111121211	1.000000
63	111121221221211111112112112	2.000000
64	111121221221211111112221111	1.000000
65	111121221221211112121111211	1.000000
66	111121221221211112121121111	1.000000
67	111121221221211122121111111	1.000000
68	111121221221211122121111211	2.000000
69	111121221221221122221111111	1.000000
70	111121221221221122221111112	1.000000
71	111121221221111112122111111	1.000000
72	111121221221112111111211211	1.000000
73	111121221221112111111221111	1.000000
74	111122121221211111121121112	1.000000
75	112112121222221122221121111	1.000000
76	112112221211211122121111211	1.000000
77	112121121212211111112112112	1.000000
78	112121121212212111111121112	1.000000
79	112121121221221122221112112	1.000000
80	121111121211211111111111111	1.000000
81	121111121211211111111111211	1.000000
82	121111121211211111111121112	1.000000
83	121111121211211111111121211	1.000000
84	121111121211211111112111111	1.000000
85	121111121221112111111211211	1.000000
86	121112121211211111111121111	1.000000
87	121112121211211122121111211	1.000000
88	121112121211211122121112211	1.000000
89	121112121221211111112121111	1.000000
90	121112122221211122121121112	1.000000
91	121121121221221122221112211	1.000000
92	121121221211211122121112112	1.000000
93	121121221221211111112112211	1.000000
94	121121221221221122221111112	1.000000
95	121211121221212111111111211	1.000000
96	211111121211211111111112111	1.000000
97	211111121211211111111121111	1.000000
98	211111121211211112111221111	1.000000
99	211111121211211112111221211	1.000000

100	21111112121121112111221221	1.000000
101	21111112121111211111211211	1.000000
102	21111112121111211111221121	1.000000
103	21111112122221112121111112	1.000000
104	21111112122121111111111111	1.000000
105	21111112122121111111121111	1.000000
106	21111112122121111111121211	1.000000
107	21111112122121112122111221	1.000000
108	21111112122111211111111112	1.000000
109	21111212121221111112211211	1.000000
110	21111212121111211111211111	1.000000
111	21111212122121122121111211	1.000000
112	21112112121221111112111111	1.000000
113	21112112121221111112121121	1.000000
114	21112112121211111111111112	1.000000
115	21112112121211211111111211	1.000000
116	21112112121211211111121211	1.000000
117	21112112121211211111221111	1.000000
118	21112112121121111111111112	1.000000
119	21112112121121112121121111	1.000000
120	21112112121121112122121112	1.000000
121	21112112121111211111111211	1.000000
122	21112112122121111111111112	1.000000
123	21112112122121111111111211	1.000000
124	21112112122121111111121112	1.000000
125	21112122122121211111111221	1.000000
126	21112122221221111112112112	1.000000
127	22111112121121111112111211	1.000000
128	22111112121111211111221111	1.000000
129	22111112122121111112111112	1.000000
130	22111212121121122121111211	1.000000
131	22111212121121122121112112	1.000000
132	22111212121121122121121211	2.000000
133	22111212121111211111211112	2.000000
134	22111212121111211111211221	1.000000
135	22111212121111211111212111	1.000000
136	22111212121111211111212112	1.000000
137	22111212122221122121112112	1.000000
138	22111212122121111111121211	1.000000
139	22111212122122122221111111	1.000000
140	22111212222121112121121111	2.000000
141	22111212222121122121111112	1.000000
142	22111212222121122121111211	1.000000
143	22111212222121122121111212	1.000000
144	22111212222121122121121112	1.000000
145	22112112121121111112111211	1.000000
146	22112112121121112121111112	1.000000
147	22112112121121112121111211	1.000000
148	22112112121121122121112211	1.000000
149	22112112121111211111111111	1.000000
150	22112112121111211111121111	1.000000
151	22112112121111211111121211	1.000000

152	22112112121111211111211211	1.000000
153	22112122121222112221112211	1.000000
154	22112122121121111111121211	1.000000
155	22112122121121122121111111	1.000000
156	22112122121121122121111211	1.000000
157	22112122121121211111211211	1.000000
158	22112122121111211111121112	1.000000
159	22112122122221122121112112	1.000000
160	22112122122121111111121211	1.000000
161	22112122122121211111111211	1.000000
162	22112122122122112221111111	1.000000
163	22112122122122112221111211	1.000000
164	22121112121121111111111112	1.000000
165	22121112122221211111111211	1.000000
166	22121112122121111111112111	2.000000

Appendix 3: Programming Scripts.

Script 1 : LDbasedGD.pl

```
#! /usr/bin/perl
use strict;
use warnings;
use DBI;
use datapro;



=head
this is a script to calculate the likelihood of two populations being
genetically distinct, by using the LD information of the genotypic data
firstly two random groups are sampled from the pooled combined sample,
then for each of the two random groups, Run the EM algorithm for a list of
rsnumbers using genotypes from a file. The markers in the file are
considered to be
a single dataset, where the EM algorithm is run for each marker compared
to all other markers.
then get the correlation value for all the pairwise LD relationships.
this process is repeated an x number of times as specified in the command
line during run time.
the the distribution of the difference in correlation values for all the
permutaions is drawn, and the probability of observing the real data is
calculated from it's place in the distribution.

=head2 arguments to script
invoke example: LDbasedGD.pl inputfile 100 98
[0]: input file (one rsnumber per line, first field is rsnumber, next
field is space-separated string of genotypes 11 12 12 22 11 11)
[1]: is the number of individuals in the first population group
[2]:  number iterations for random sampling of two groups from the pooled
individuals.

=head3
the method used for generating random samples from the data is as follows:
Suppose, we have an empty list. We pick a random number between 1 and
10^12 and add it to the list only if it was not already picked before,
i.e. if it is not already contained in the list.
We then do the same thing again and again until we have eventually
collected 10^6 distinct numbers.
Now we sort the set ascending and return it.

=cut
my $st = time;
my $now = gmtime;
print "Starting process at $now...\n";



die "Invoke with input filename, number of individuals in the first group
and number of iterations\n" unless @ARGV == 3 ;
my $infile = $ARGV[0];
my $first_group = $ARGV[1];
```

```
my $iterations = $ARGV[2];

my $outfile ; # = "twogroups_LDvalues_10000.txt";

if ($ARGV[0] =~ /(\w+)\.txt/) {
          $outfile = $1."_LDvalues_JN".$iterations.".txt";

              }


print "Opening input file $infile...";
open FH, $infile or die "Could not open infile $infile\n";
print "done\n";

my @G = <FH>;
my $g = $G[0];
chomp $g;
my @H = split /\s+/, $g;

my $totalindivs = (scalar @H -1);

my @markers;
my @details;
my $markernum = 1;
my @firstgroup;
my @secondgroup;
my $lines = 0;
foreach my $ln (@G) {
   $lines++;

   chomp $ln;

   my @indivs = split /\s+/, $ln;
   my $rs = shift @indivs;
  #my $totalindivs = scalar @indivs;
  #print "total num num num of indivs is $totalindivs in this
instance.????\n";

   my $sub_record;
   my @this_record;
   my $dd = 0;
   while ($dd < $first_group) {

   my $ra = $indivs[$dd];
   push @this_record , $ra;

$dd++;
   }
   #print "Retrieved suitable genotypes for ", $dd, " for the first
group.\n";

$sub_record = join " ", @this_record;
push @firstgroup, $sub_record;

my $sub_record2;
my @this_record2;
   for ( my $ss = $first_group; $ss < $totalindivs;$ss++ ) {

my $amon = $indivs[$ss];

push @this_record2 , $amon;
```

```
    }

    $sub_record2 = join " ", @this_record2;
    push @secondgroup, $sub_record2;


      push @markers, $rs;
    push @details, [$markernum,$rs];

    $markernum++;

}

#now we have two arrays with the same rs genotyping records, each array
represents one of our real groups of individuals.

close FH;
#print "Retrieved suitable genotypes for ", $dd, " for the first group and
", scalar @secondgroup, " for the second group of $lines markers.\n";

print "Preparing to calculate metrics for the two real groups.\n";


my @LD_firstgroup;
for (my $p=0; $p < scalar @firstgroup-1; $p++) {

   for (my $q=$p+1; $q < scalar @firstgroup; $q++) {

      my $rs1r = $markers[$p];
      my $rs2r = $markers[$q];
      my $geno_rs1r = $firstgroup[$p]; # string of 11 12 11 11 12 22 00 11
     #print $geno_rs1r ;
      my $geno_rs2r = $firstgroup[$q]; # string of 11 12 11 11 12 22 00 11
    #print $geno_rs2r ;
      my $rs1_in_dset1r = $p+1;
      my $rs2_in_dset1r = $q+1;

      my ($maj1r,$min1r,$maj2r,$min2r,$f11r,$f12r,$f21r,$f22r) =
datapro::EM_algorithm($geno_rs1r,$geno_rs2r);

      # check that the EM algorithm gave valid results
      unless (defined $maj1r and defined $min1r and defined $maj2r and
defined $min2r and defined $f11r and defined $f12r and defined $f21r and
defined $f22r) {
         die "ERROR: EM algorithm for:\n$geno_rs1r\nvs\n$geno_rs2r\n gave
maj1 = $maj1r, min1 = $min1r, maj2 = $maj2r, min2 = $min2r, f11 = $f11r,
f12 = $f12r, f21 = $f21r, f22 = $f22r\n";
      }

      # calc D using the PAIRWISE allele frequencies from the EM algorithm
      my ($Dr,$absDr,$Dprimer,$absDprimer) =
datapro::calc_D($maj1r,$min1r,$maj2r,$min2r,$f11r,$f12r,$f21r,$f22r);

      # check that calc_D gave valid results
      unless (defined $Dr and defined $absDr and defined $Dprimer and
defined $absDprimer) {
         die "ERROR: calc_D gave D = $Dr, absD = $absDr, Dprime =
$Dprimer, absDprime = $absDprimer\n";
      }
```

```
      # calc dalta2 using the PAIRWISE allele frequencies from the EM
algorithm
      my $delta2r =
datapro::calc_delta2($maj1r,$min1r,$maj2r,$min2r,$f11r,$f12r,$f21r,$f22r);

      # check that calc_delta2 gave valid results
      unless (defined $delta2r) {
         die "ERROR: calc_delta2 gave delta2 = $delta2r\n";
      }

      my $pairwiser = join
"\t",$rs1_in_dset1r,$rs2_in_dset1r,$rs1r,$rs2r,$min1r,$min2r,$absDprimer,$
delta2r;
      push @LD_firstgroup, $pairwiser;

   }
}

my @LD_secondgroup;
for (my $u =0; $u < scalar @secondgroup-1; $u++) {
   for (my $v=$u+1; $v < scalar @secondgroup; $v++) {

      my $rrs1r = $markers[$u];
      my $rrs2r = $markers[$v];
      my $geno_rrs1r = $secondgroup[$u]; # string of 11 12 11 11 12 22 00
11
     #print $geno_rs1r ;
      my $geno_rrs2r = $secondgroup[$v]; # string of 11 12 11 11 12 22 00
11
    #print $geno_rs2r ;
      my $rs1_in_dset2r = $u+1;
      my $rs2_in_dset2r = $v+1;

      my ($mmaj1r,$mmin1r,$mmaj2r,$mmin2r,$ff11r,$ff12r,$ff21r,$ff22r) =
datapro::EM_algorithm($geno_rrs1r,$geno_rrs2r);

      # check that the EM algorithm gave valid results
      unless (defined $mmaj1r and defined $mmin1r and defined $mmaj2r and
defined $mmin2r and defined $ff11r and defined $ff12r and defined $ff21r
and defined $ff22r) {
         die "ERROR: EM algorithm for:\n$geno_rrs1r\nvs\n$geno_rrs2r\n
gave maj1 = $mmaj1r, min1 = $mmin1r, maj2 = $mmaj2r, min2 = $mmin2r, f11 =
$ff11r, f12 = $ff12r, f21 = $ff21r, f22 = $ff22r\n";
      }

      # calc D using the PAIRWISE allele frequencies from the EM algorithm
      my ($D2r,$absD2r,$Dprime2r,$absDprime2r) =
datapro::calc_D($mmaj1r,$mmin1r,$mmaj2r,$mmin2r,$ff11r,$ff12r,$ff21r,$ff22
r);

      # check that calc_D gave valid results
      unless (defined $D2r and defined $absD2r and defined $Dprime2r and
defined $absDprime2r) {
         die "ERROR: calc_D gave D = $D2r, absD = $absD2r, Dprime =
$Dprime2r, absDprime = $absDprime2r\n";
      }

      # calc dalta2 using the PAIRWISE allele frequencies from the EM
algorithm
```

```
      my $delta22r =
datapro::calc_delta2($mmaj1r,$mmin1r,$mmaj2r,$mmin2r,$ff11r,$ff12r,$ff21r,
$ff22r);

      # check that calc_delta2 gave valid results
      unless (defined $delta22r) {
         die "ERROR: calc_delta2 gave delta2 = $delta22r\n";
      }

      my $pairwise2r = join
"\t",$rs1_in_dset2r,$rs2_in_dset2r,$rrs1r,$rrs2r,$mmin1r,$mmin2r,$absDprim
e2r,$delta22r;
      push @LD_secondgroup, $pairwise2r;

   }
}


print "done calculating the LD values for the two real groups.\n";

open FH_OUT, ">$outfile" or die "Could not open outfile $outfile for
writing\n";

print "output file opened for writing results of real groups .\n";

foreach my $pair_record1 (@LD_firstgroup) {
    my @aa = split "\t",$pair_record1;
   my $rs1_1 = $aa[2];
   my $rs2_1 = $aa[3];
   my $comb_MAF1 = $aa[4] + $aa[5];
foreach my $pair_record2 (@LD_secondgroup) {
   my @bb = split "\t",$pair_record2;
   my $rs1_2 = $bb[2];
   my $rs2_2 = $bb[3];
my $comb_MAF2 = $bb[4] + $bb[5];

if (($rs1_1 eq $rs1_2 or $rs1_1 eq $rs2_2) and ($rs2_1 eq $rs1_2 or $rs2_1
eq $rs2_2)) {


my $r_LD = join "\t",($aa[7],$bb[7]); #,$comb_MAF1,$comb_MAF2);
print FH_OUT "$r_LD\n";

}
}
}

close FH_OUT;


#system "nedit $outfile &";
# now we have a file with two columns, the first column is the delta2
values of the first real group, the second is the matched values for the
second group.

#next we want to read the file with the two real groups' LD values with R
and get Spearman's Rank Correlation Coefficient for them.

my $var = `R --vanilla <KOKA_SAL.R> logfile`;
```

```
my $o = 0;
while ($o < $iterations) {

# to generate random lists: The rand() function is used to generate random
numbers.
# By default it generates a number between 0 and 1, however you can pass
it a maximum and it will generate numbers between 0 and that number.


my %random_group;
my @K;
my $k = 0;

my $c =1;
while ($k < $first_group) {
  my $random_number = int(rand($totalindivs)) + 1; #This gives you an
integer from 1 to numberofindivs inclusive:
  unless ( exists $random_group{$random_number}) {

$random_group{$random_number} = $c;

$c++;
@K = keys %random_group;
$k = scalar @K;
}
}

# now we have a randomly generated list out of the total pooled
individuals, equal in number to the first pop group.


my @second_group;
for (my $f=1; $f <= $totalindivs; $f++) {
unless (exists $random_group{$f}) {
push @second_group, $f;
}
}

#now we have a second list with the remaining numbers.

print "Processing markers in input file to pick genotypes of random
groups...\n";

my $markernum = 1;
my @new_firstgroup;
my @new_secondgroup;
foreach my $line (@G) {


chomp $line;

my @fields = split /\s+/, $line;
my $rs = shift @fields;


my $new_record;
my @some_record;
#push @some_record, $rs;
foreach my $d (@K) {
```

```
push @some_record , $fields[$d];
}

$new_record = join " ", @some_record;
push @new_firstgroup, $new_record;

my $new_record2;
my @some_record2;
#push @some_record2, $rs;
foreach my $s (@second_group) {

push @some_record2 , $fields[$s];
}

$new_record2 = join " ", @some_record2;
   push @new_secondgroup, $new_record2;

}

#now we have two arrays with the same rs genotyping records, each array
represents a different group of individuals.


print "Preparing to calculate metrics for the two random groups.\n";

#my $processed = 0;


my @LDvalues_firstgroup;
for (my $i=0; $i < scalar @new_firstgroup-1; $i++) {
   for (my $j=$i+1; $j < scalar @new_firstgroup; $j++) {

      my $rs1 = $markers[$i];
      my $rs2 = $markers[$j];
      my $geno_rs1 = $new_firstgroup[$i]; # string of 11 12 11 11 12 22 00
11
     #print $geno_rs1 ;
      my $geno_rs2 = $new_firstgroup[$j]; # string of 11 12 11 11 12 22 00
11
    #print $geno_rs2 ;
      my $rs1_in_dset1 = $i+1;
      my $rs2_in_dset1 = $j+1;

      my ($maj1,$min1,$maj2,$min2,$f11,$f12,$f21,$f22) =
datapro::EM_algorithm($geno_rs1,$geno_rs2);

      # check that the EM algorithm gave valid results
      unless (defined $maj1 and defined $min1 and defined $maj2 and
defined $min2 and defined $f11 and defined $f12 and defined $f21 and
defined $f22) {
         die "ERROR: EM algorithm for:\n$geno_rs1\nvs\n$geno_rs2\n gave
maj1 = $maj1, min1 = $min1, maj2 = $maj2, min2 = $min2, f11 = $f11, f12 =
$f12, f21 = $f21, f22 = $f22\n";
      }

      # calc D using the PAIRWISE allele frequencies from the EM algorithm
      my ($D,$absD,$Dprime,$absDprime) =
datapro::calc_D($maj1,$min1,$maj2,$min2,$f11,$f12,$f21,$f22);

      # check that calc_D gave valid results
```

```
      unless (defined $D and defined $absD and defined $Dprime and defined
$absDprime) {
         die "ERROR: calc_D gave D = $D, absD = $absD, Dprime = $Dprime,
absDprime = $absDprime\n";
      }

      # calc dalta2 using the PAIRWISE allele frequencies from the EM
algorithm
      my $delta2 =
datapro::calc_delta2($maj1,$min1,$maj2,$min2,$f11,$f12,$f21,$f22);

      # check that calc_delta2 gave valid results
      unless (defined $delta2) {
         die "ERROR: calc_delta2 gave delta2 = $delta2\n";
      }

      my $pairwise = join
"\t",$rs1_in_dset1,$rs2_in_dset1,$rs1,$rs2,$min1,$min2,$absDprime,$delta2;
      push @LDvalues_firstgroup, $pairwise;

   }
}

my @LDvalues_secondgroup;
for (my $n=0; $n < scalar @new_secondgroup-1; $n++) {
   for (my $m=$n+1; $m < scalar @new_secondgroup; $m++) {

      my $rrs1 = $markers[$n];
      my $rrs2 = $markers[$m];
      my $geno_rrs1 = $new_secondgroup[$n]; # string of 11 12 11 11 12 22
00 11
     #print $geno_rs1 ;
      my $geno_rrs2 = $new_secondgroup[$m]; # string of 11 12 11 11 12 22
00 11
    #print $geno_rs2 ;
      my $rs1_in_dset2 = $n+1;
      my $rs2_in_dset2 = $m+1;

      my ($mmaj1,$mmin1,$mmaj2,$mmin2,$ff11,$ff12,$ff21,$ff22) =
datapro::EM_algorithm($geno_rrs1,$geno_rrs2);

      # check that the EM algorithm gave valid results
      unless (defined $mmaj1 and defined $mmin1 and defined $mmaj2 and
defined $mmin2 and defined $ff11 and defined $ff12 and defined $ff21 and
defined $ff22) {
         die "ERROR: EM algorithm for:\n$geno_rrs1\nvs\n$geno_rrs2\n gave
maj1 = $mmaj1, min1 = $mmin1, maj2 = $mmaj2, min2 = $mmin2, f11 = $ff11,
f12 = $ff12, f21 = $ff21, f22 = $ff22\n";
      }

      # calc D using the PAIRWISE allele frequencies from the EM algorithm
      my ($D2,$absD2,$Dprime2,$absDprime2) =
datapro::calc_D($mmaj1,$mmin1,$mmaj2,$mmin2,$ff11,$ff12,$ff21,$ff22);

      # check that calc_D gave valid results
      unless (defined $D2 and defined $absD2 and defined $Dprime2 and
defined $absDprime2) {
         die "ERROR: calc_D gave D = $D2, absD = $absD2, Dprime =
$Dprime2, absDprime = $absDprime2\n";
      }
```

```
      # calc dalta2 using the PAIRWISE allele frequencies from the EM
algorithm
      my $delta22 =
datapro::calc_delta2($mmaj1,$mmin1,$mmaj2,$mmin2,$ff11,$ff12,$ff21,$ff22);

      # check that calc_delta2 gave valid results
      unless (defined $delta22) {
         die "ERROR: calc_delta2 gave delta2 = $delta22\n";
      }

      my $pairwise2 = join
"\t",$rs1_in_dset2,$rs2_in_dset2,$rrs1,$rrs2,$mmin1,$mmin2,$absDprime2,$de
lta22;
      push @LDvalues_secondgroup, $pairwise2;
   }
}


open FH_OUT, ">$outfile" or die "Could not open outfile $outfile for
writing\n";

print "opening output file to write results of the random groups.\n";

foreach my $pair_1 (@LDvalues_firstgroup) {
    my @aaa = split "\t",$pair_1;
   my $rrs1_1 = $aaa[2];
   my $rrs2_1 = $aaa[3];
   my $com_MAF1 = $aaa[4] + $aaa[5];
foreach my $pair_2 (@LDvalues_secondgroup) {
   my @bbb = split "\t",$pair_2;
   my $rrs1_2 = $bbb[2];
   my $rrs2_2 = $bbb[3];
my $com_MAF2 = $bbb[4] + $bbb[5];

if (($rrs1_1 eq $rrs1_2 or $rrs1_1 eq $rrs2_2) and ($rrs2_1 eq $rrs1_2 or
$rrs2_1 eq $rrs2_2)) {


my $twogroup_LD = join "\t",($aaa[7],$bbb[7]); #,$com_MAF1,$com_MAF2);
print FH_OUT "$twogroup_LD\n";

}
}
}

close FH_OUT;
print "done calculating LD values for permutation number $o\n";
# now we have a file with two columns, the first column is the delta2
values of the first random group, the second is the matched values for the
second group.

#next we want to read the file with the two random groups' LD values with
R and get Spearman's Rank Correlation Coefficient for them.

my $var = `R --vanilla <KOKA_SAL.R> logfile`;



$o++;
}
```

```
#all the permutations' correlation values are stored in the file ex.data,
the first value should be that of the real data.


=cut
=head
#display it by plotting a graph
#get the probability of observing the real data.
#my $variable = `R --vanilla <Spearman.R> logfile`;

my $infile2 = "ex_10000.data";
open FH_IN, $infile2 or die "Could not open infile $infile2\n";


my @file_lines = <FH_IN>;
my $total_lines = (scalar @file_lines);


my $outfile2 = "prob_values_10000.txt";
open FH_DIS, ">$outfile2" or die "Could not open outfile $outfile2 for
writing\n";

my @result;
for (my $w = 9; $w <= $total_lines; $w = $w + 10) {

 my $y = $w +10;
 my $cor_LD = $file_lines[$w];
 my $cor_maf = $file_lines[$y];
 my $diff = $cor_maf - $cor_LD;
 push @result, $diff;
 print FH_DIS "$cor_LD\n";

 }


close FH_IN;
close FH_DIS;

=cut

my $dur = time - $st;
my $finish = gmtime;
print "Finished after $dur seconds. Results in $outfile\nTime at finish:
$finish\n";
```

Script 2: ext_hap_freq.pl. Script used for analysis of the HapMap data for extended high frequency haplotypes (chapter 5).

```
#!/usr/bin/perl
use strict;
use warnings;

=head1 ext_hap_freq.pl

Slides windows of chosen size across haplotypes supplied.
Outputs (for each position of the window), the numbers of identical
haplotypes, in descending order.
A typical line could look like:
7, 3, 2, 2
which means that that window (i.e. window of fixed size but starting at
that position) had 7 haplotypes
the same (throughout the window), and also 3 haplotypes the same, then 2,
then another 2, and the rest were distinct.
However if all the haplotypes are distinct, we output
1
as a single line, so that this (very common) case is countable.  (We don't
bother to append 1s to the end of
lines where there are identical haplotypes.)
This scripts also calculates the average recombination rate for each
sliding window using a file of estimated recombination rates downloaded
from HapMap. The filename is inferred from the hapfile name.

=head2 Arguments to script

 Invoke as perl ext_hap_freq.pl hapfile 75 1
 to process the haplotypes in file hapfile with a window size of 75 and a
window shift of 1

=head2 Haplotype file format

 Each line should be a single haplotype only.  It does not matter whether
the numbers/characters are separated by spaces.
 Each line should have the same number of (non-space) characters on it.

Example

1211112121212211212121212121212
2121112122121221211212121221221
...


=cut

# where are the haplotype files?
my $REFDIR = "./";

# check command line arguments
die "Invoke as 'perl ext_hap_freq.pl hapfile 75 1' to read hapfile, window
size 75 with window shift 1.\n"
   unless 3 == scalar @ARGV and $ARGV[1]>0 and $ARGV[2]>0 and
$ARGV[2]<=$ARGV[1];

my $infile = $REFDIR . $ARGV[0];
my $windowsize = $ARGV[1];
```

```
my $windowshift = $ARGV[2];
my $legend_file = "";
my $recom_file = "";
if ($ARGV[0] =~ /(\w+(chr\d+|chr[XY])_\w+)\.phased/) {
   $legend_file = $REFDIR . $1 . "_legend.txt";
   $recom_file = $REFDIR . "recomb_rate_" . $2 . ".txt";
}
die "Legend file $legend_file not found.\n" unless -f $legend_file;
die "Recombination rate file $recom_file not found.\n" unless -f
$recom_file;

print "\nusing $infile\nlegend file = $legend_file\nrecom file =
$recom_file\n";

# output files
my $hapfile = "haps.out";
my $freqfile = "freqs.out";
my $chkfile = "long.out";
my $posfile = "pos.out";
my $winfile = "win.out";

my $st = time;

# all trials are recorded in the log. This is appended to each time a
trial is run
my $haplog = "haplog.txt";
die "Can't find the log file $haplog\n" unless -e $haplog;
my $lognum = `cat $haplog | wc -l`;
chomp $lognum;
$lognum =~ s/^\s+//; # remove leading whitespace
$lognum =~ s/\s+$//; # remove trailing whitespace

# copy selected output files to unique names at the end
my $hapcopy = "$lognum\_haps.txt";
my $freqcopy = "$lognum\_freqs.txt";
my $poscopy = "$lognum\_pos.txt";
my $wincopy = "$lognum\_win.txt";

# read in recombination rates file
my ($details,$starts,$stops) = read_recomb_rate_file($recom_file);

# populate array of haplotypes as strings
my @haps;

# read from haplotype file
local *FH;
open (FH, $infile) or die "Couldn't open $infile to read.\n";
my $full_haplength;
while (<FH>) {
      chomp;
      my @temp = split; # splits on whitespace
   my $thishap = join("", @temp); # haplotype as a string like
12112122112121212121
   if (defined $full_haplength) {
      die "Didn't find a haplotype of length $full_haplength" unless
length($thishap) == $full_haplength;
   } else { # infer length to check subsequent lines against from first
line
      $full_haplength = length($thishap);
   }
      push @haps, $thishap;
```

```
}
close FH;

# read in chromosome coordinates of markers from legend file
# one marker per line, in chromosomal positional order (hence same order
as haplotypes file)
# store the coordinates in @coordinates, where element 456 contains the
coordinate for the 457th marker in the haplotype
my @coordinates;
open FH_LEGEND, $legend_file or die "Could not open legend file
$legend_file.\n";
<FH_LEGEND>; # first line is header line
while (my $line = <FH_LEGEND>) {

   my ($pos,$zero,$one) = split /\s+/, $line;
   push @coordinates, $pos;

}
close FH_LEGEND;

$| = 1; # don't buffer output

my $numchr = scalar @haps;
print "Read $numchr haplotypes of length $full_haplength.\n";

# open main output file
open (FH, ">$hapfile") or die "Couldn't open $hapfile to write ";
my $old_fh = select(FH);
$| = 1; # don't buffer output to this file
select($old_fh);

# open outfile to write window coords to
open (FH_POS, ">$posfile") or die "Couldn't open $posfile to write ";

# open outfile to write window number, max haplotype frequency, average
recombination rate, average genetic distance, n_recombination windows
open (FH_WIN, ">$winfile") or die "Couldn't open $winfile to write ";
print FH_WIN "window\tmax_hap_freq\tavg_Rate\tapprox_cM\n";

# calculation

# loop through the window positions available (pos is starting position, 0
meaning starting from first SNP)
my $max_simhap = 1;
my $max_pos = 0;
my $numpos = 0;
my %numchr_to_freq; # key is number of chromosomes, value is number of
times that number of chromosomes were identical
for (my $pos = 0; $pos+$windowsize <= $full_haplength; $pos+=$windowshift)
{

   my @hapcounts = count_similar_haps(\@haps, $pos, $windowsize);

   my $markerpos = $pos+1;
   my $winstart = $coordinates[$pos];
   my $winend = $coordinates[$pos+$windowsize-1];

   my $maxhaps = 1; # the maximum number of identical haplotypes at this
position
   if (scalar @hapcounts) {
```

```
        $maxhaps = $hapcounts[0];

        if ($hapcounts[0] > $max_simhap) { # keep track of the highest
number of identical haplotypes across positions and the position
            $max_simhap = $hapcounts[0];
            $max_pos = $pos;
        }
    }

    for my $hc (@hapcounts) { # record the frequency of identical
chromosomes (>=2) at this position
        $numchr_to_freq{$hc}++;
    }

    my $outputline = (scalar @hapcounts) ? (join ",", @hapcounts) : '1';
        print FH "$outputline\n";
        $numpos++;

    my ($string,$avg_recom_rate,$approx_cM) =
calc_recombination_rate($winstart,$winend,$details,$starts,$stops);

    print FH_WIN "$numpos\t$maxhaps\t$avg_recom_rate\t$approx_cM\n";
    print FH_POS "window $numpos starts at marker $markerpos and has coords
$winstart-$winend. Rate = $avg_recom_rate cM.Mb over $approx_cM cM
($string)\n";
}

close FH;
close FH_POS;
close FH_WIN;

# print the frequencies to a file for plotting graphs
process_freqs(\%numchr_to_freq,$numchr,$numpos,$freqfile);

# extract and print to file the haplotype strings for the window that had
the highest identical haplotype frequency
extract_best_window(\@haps,$max_pos,$windowsize,$chkfile);

# print info about this run to the logfile
open FH_LOG, ">>$haplog" or die "Could not open logfile $haplog.\n";
print FH_LOG
"$lognum\t$windowsize\t$infile\t$numchr\t$full_haplength\t$max_simhap\t$ma
x_pos\t$windowshift\t$numpos\n";
close FH_LOG;

print "trial $lognum. Window = $windowsize markers. $numchr haplotypes of
$full_haplength markers ($numpos window positions and windowshift =
$windowshift). $max_simhap of ", scalar @haps, " chromosomes were
identical at position $max_pos\n";
print "See results:\n";
print "$hapfile - frequencies of haplotypes at each position\n";
print "$freqfile - overall frequencies of haplotypes across all
positions\n";
print "$posfile - window positions and associated recombination rates\n";
print "$winfile - window number, max number of identical haplotypes,
average recombination rates and genetic distance of window\n";
print "Finished in ", time-$st, " seconds. Log information is in
$haplog.\n";
system "cp $hapfile $hapcopy";
system "cp $freqfile $freqcopy";
```

```
system "cp $posfile $poscopy";
system "cp $winfile $wincopy";
system "nedit $hapfile $freqfile $posfile $winfile &";


=head1 count_similar_haps

For a particular window position, examine the substring of haplotypes of a
given size and
count the numbers of identical haplotypes.  We do not include distinct
haplotypes.

=head2 Arguments

 $_[0] : reference to array of haplotypes
 $_[1] : start position (0-based: 0 means beginning of string)
 $_[2] : size of window

Returns empty list if all haplotypes are distinct in this region, or a
list of numbers
like 7, 3, 2, 2 (ordered descending) which represents the number of
identical hits.

=cut

sub count_similar_haps {

   my ($ref_haps, $start, $size) = @_;

   my %subhaps; # hash with key of subhaplotype string, value number of
occurrences

      for my $full_hap (@{$ref_haps}) {
            my $sub_hap = substr($full_hap, $start, $size);
         $subhaps{$sub_hap}++; # will be 1 if not encountered before,
otherwise increment
   }


   # now find those that are not 1
   my @interesting;
   for (values %subhaps) {
      push @interesting, $_ unless $_ == 1;
   }

   return sort {$b<=>$a} @interesting;

} # end sub count_similar_haps

=head1 process_freqs

Calculate the frequency of chromosomes being identical for all number of
chromosomes between 2 and $numchr totalled over all positions.

=head2 arguments

 [0] : a hash of number of chromosomes => freq of being identical
 [1] : total number of chromosomes examined
 [2] : total number of windows used
 [3] : file to write to
```

```
=cut

sub process_freqs {

   my ($numchr_to_freq,$numchr,$numpos,$freqfile) = @_;

   open FH, ">$freqfile" or die "Could not open $freqfile.\n";
   for (my $i = 2; $i <= $numchr; $i++) {
      print FH "$i\t";
   }
   print FH "\n";

   for (my $i = 2; $i <= $numchr; $i++) {

      if (defined $$numchr_to_freq{$i}) {

         # this is the number of times that $i of $numchr chromosomes are
identical
         my $freq = $$numchr_to_freq{$i};
         print FH "$freq\t";

      } else {

         print FH "0\t";
      }

   }
   print FH "\n";
   close FH;

} # end sub process_freqs

=head1 extract_best_window

Extract the haplotype strings for the window that gave the greatest number
of identical haplotypes

=head2 Arguments

 [0] : ref to array of haplotype strings (one element per chromosome)
 [1] : the window position that gives the highest number of identical
haplotypes
 [2] : the size of the window

=cut

sub extract_best_window {

   my ($ref_haps,$max_pos,$windowsize,$chkfile) = @_;

my @temp;
for my $full_hap (@{$ref_haps}) {
         my $sub_hap = substr($full_hap, $max_pos, $windowsize);
   push @temp, $sub_hap;
}

my @sorted = sort(@temp);

open FH, ">$chkfile" or die "Could not open $chkfile.\n";
for my $h (@sorted) {
   print FH "$h\n";
```

```
    }
    close FH;

} # end sub extract_best_window

=head1 read_recomb_rate_file

Read in a recombination rates file from HapMap. The following fields are
required:

chrom start stop Rate_cM.Mb Avg_cM Gen_map_cM

It is assumed that the windows are in ascending chromosomal coordinate
order

=head2 Arguments

 $_[0]: filename of input file

 returns a ref to array of ordered recom window start coords, an array of
ordered recom window end coords, and an array of recom window details
 in same order as start and end coords

=cut

sub read_recomb_rate_file {

   my $filename = $_[0];

   my @details;
   my @starts;
   my @stops;

   open FH, $filename or die "Could not open input file $filename\n";

   <FH>; # skip the first line (header line)
   while (my $line = <FH>) {

      chomp $line;

      my ($chrom,$start,$stop,$rate,$Avg_cM,$Gen_map_cM) = split /\s+/,
$line;

      push @details, "$rate\t$Avg_cM\t$Gen_map_cM";
      push @starts, $start;
      push @stops, $stop;

   }
   close FH;

   return (\@details,\@starts,\@stops);

} # end sub read_recomb_rate_file


sub calc_recombination_rate {

   my ($winstart,$winend,$ref_details,$ref_starts,$ref_stops) = @_;

   my $winstart_index = find_max_coord($winstart, $ref_starts);
   my $winend_index = find_max_coord($winend, $ref_starts);
```

```
   my $total_recom_windows = 0;
   my $total_rate = 0;
   my $total_cM = 0;
   for my $ind ($winstart_index .. $winend_index) {

      my $details = $$ref_details[$ind];
      my ($rate,$Avg_cM,$Gen_map_cM) = split /\t+/, $details;

      $total_rate += $rate;
      $total_cM += $Avg_cM;
      #print "rate = $rate, total rate = $total_rate, cM = $Avg_cM, total
cM = $total_cM\n";

      $total_recom_windows++;

   }

   my $avg_rate = $total_rate/$total_recom_windows;
   #my $avg_cM = $total_cM/$total_recom_windows;

   #print "TOTAL RATE = $total_rate, avg rate = $avg_rate\n";
   #print "TOTAL cM = $total_cM, avg cM = $avg_cM\n";

   my $string = "winstart in recom window at index $winstart_index, winend
at $winend_index (total = $total_recom_windows)";

   return ($string,$avg_rate,$total_cM);

} # end sub read_recombination_rate

=head1 find_max_coord

Search through array of genes to find the maximum index of genes whose
start coord is less than/equal to the snp

=head2 Arguments

 [0] : coordinate of marker
 [1] : reference to array of arrays sorted in chromosome start order
         (routine is likely to fail if the array is empty)

Returns the appropriate index

=cut

sub find_max_coord {

      my ($coord, $arr) = @_;

      my $ind; # index of $arr corresponding to the start coordinate of a
boundary gene

      # binary search
      my $max = (scalar @{$arr}) - 1;
      $ind = int ($max / 2); # current position
      my $up = $max; # top of current window we are searching in
      my $down = 0; # bottom of current window we are searching in
      my $got_flag = 0;
   my %already_checked; # to keep a record of indices we've already
checked
```

```
        while (!$got_flag) {

                last if defined $already_checked{$ind} and
$already_checked{$ind} == 1;
                $already_checked{$ind} = 1;

        if ($ind>0 and $ind<$max) { # normal case

           # start by checking whether the current position is OK
                # note that if we get to the edges the behaviour is
different
                my $gene_coord = $arr->[$ind];
           my $gene_coord_plus1 = $arr->[$ind+1];

           # now assess which of three possibilities applies:
           # possibility 1: this is the right place. The snp coord lies
between gene_coord and gene_coord_plus1
           if ($coord >= $gene_coord and $coord < $gene_coord_plus1) {
              $got_flag = 1;

                   # possibility 2: both gene_coord and gene_coord_plus1
are lower than the snp coord. move to higher coords
           } elsif ($coord > $gene_coord and $coord >= $gene_coord_plus1) {

              # don't do the halving if total window size is small
                        if (($up-$down)>=4) {
                 $down = $ind;
                              $ind = int(($up-$down)/2) + $down;
                        } else { # just walk right
                 $ind++;
                        }
                        $ind = $max if ($ind>$max);

           # possibility 3: both gene_coord and gene_coord_plus1 are higher
than the snp coord. move to lower coords
           } elsif ($coord < $gene_coord and $coord < $gene_coord_plus1) {

              # don't do the halving if total window size is small
                        if (($up-$down)>=4) {
                 $up = $ind;
                              $ind = int(($up-$down)/2) + $down;
                        } else { # just walk left
                 $ind--;
                        }
              $ind = 0 if ($ind<0);

                 }

              } else {
           # $ind is maximum or zero, we must stop the loop
              # snp_coord may be less than all gene starts on this chrom
              $got_flag = 1;
                 }

        } # end of while !$got_flag

        return $ind;

} # end sub find_max_coord
```

```
=head1 find_min_coord

Search through array of genes to find the minimum index of genes whose end
coord is greater than/equal to the snp

=head2 Arguments

 [0] : coordinate of marker
 [1] : reference to array of arrays sorted in chromosome end order
          (routine is likely to fail if the array is empty)

Returns the appropriate index

=cut

sub find_min_coord {

      my ($coord, $arr) = @_;

   my $ind; # index of $arr corresponding to the start/end coordinate of a
boundary gene

   # binary search
      my $max = (scalar @{$arr}) - 1;
      $ind = int ($max / 2); # current position
      my $up = $max; # top of current window we are searching in
      my $down = 0; # bottom of current window we are searching in
      my $got_flag = 0;
   my %already_checked; # to keep a record of indices we've already
checked

   while (!$got_flag) {
            last if defined $already_checked{$ind} and
$already_checked{$ind} == 1;
            $already_checked{$ind} = 1;

      if ($ind>0 and $ind<$max) { # normal case

         # start by checking whether the current position is OK
               # note that if we get to the edges the behaviour is
different
               my $gene_coord = $arr->[$ind][1];
         my $gene_coord_min1 = $arr->[$ind-1][1];

         # now assess which of three possibilities applies:
         # possibility 1: this is the right place. The snp coord lies
between gene_coord and gene_coord_min1
         if ($coord <= $gene_coord and $coord > $gene_coord_min1) {
            $got_flag = 1;
                  # possibility 3: both gene_coord and gene_coord_min1 are
lower than the snp coord. move to higher coords
         } elsif ($coord > $gene_coord and $coord > $gene_coord_min1) {

            # don't do the halving if total window size is small
                     if (($up-$down)>=4) {
               $down = $ind;
                           $ind = int(($up-$down)/2) + $down;
                     } else { # just walk right
               $ind++;
```

```
                        }
                        $ind = $max if ($ind>$max);

        # possibility 3: both gene_coord and gene_coord_min1 are higher
than the snp coord. move to lower coords
        } elsif ($coord < $gene_coord and $coord <= $gene_coord_min1) {

           # don't do the halving if total window size is small
                        if (($up-$down)>=4) {
              $up = $ind;
                              $ind = int(($up-$down)/2) + $down;
                        } else { # just walk left
              $ind--;
                        }
           $ind = 0 if ($ind<0);

                     }

                        } else {
                     # $ind is maximum or zero, we must stop the loop
                     # snp_coord may be greater than all gene ends on this chrom
                     $got_flag = 1;
                        }

        } # end of while !$got_flag

        return $ind;

} # end sub find_min_coord
```

Script 3: exthap-TDT-cm.pl. Script used for analysis of the MalariaGen Gambian trio case-control data for extended high frequency haplotypes (chapter 6).

```
#!/usr/bin/perl
use strict;
use warnings;

=head1 exthap-TDT-cm.pl

Slides windows of a fixed chosen genetic distance across haplotypes
supplied.
Outputs:
hapfrequency file: (for each position of the window), the numbers of
identical haplotypes, in descending order.
A typical line could look like:
7, 3, 2, 2
which means that that window had 7 haplotypes the same (throughout the
window), and also 3 haplotypes the same, then 2, then another 2, and the
rest were distinct.
However if all the haplotypes are distinct, we output 1
as a single line, so that this (very common) case is countable.  (We don't
bother to append 1s to the end of
lines where there are identical haplotypes.)
windows file: details of each sliding window start and stop position, size
in physical and genetic distance.
summary file: with number of highest hap frequency across all windows, and
window position

=head2 Arguments to script

 Invoke as perl exthap-affy-cm.pl hapfile 1 0.5
 to process the haplotypes in file hapfile with a window size of 1 CM and
a window shift of 0.5 CM

=head2
 Haplotype file format:

a line of individuals ID followed by two lines of phased haplotypes for
this individual
 It does not matter whether the numbers/characters are separated by
spaces.
 Each line should have the same number of (non-space) characters on it.

Example

individualID
1211112121212212121212121212121
2121112122121221211212121221221
...

legend file format:  chr      rs     cM    position

=cut

# where are the haplotype files?
my $REFDIR = "./";

# check command line arguments
die "Invoke as 'perl exthap-TDT-cm2.pl hapfile legendfile 1 0.5' to read
hapfile, window size 1 CM with window shift 0.5 CM.\n"
```

```
   unless 4 == scalar @ARGV and $ARGV[2]>0 and $ARGV[3]>0 and
$ARGV[3]<=$ARGV[2];

my $infile = $REFDIR . $ARGV[0];
my $legend_file = $REFDIR . $ARGV[1];
my $winsize = $ARGV[2];
my $winshift = $ARGV[3];


print "\nusing $infile\nlegend file = $legend_file\n";

# output files
my $hapfile = "haps.out";
my $freqfile = "freqs.out";
my $chkfile = "long.out";
my $posfile = "pos.out";
my $winfile = "win.out";

my $st = time;

# all trials are recorded in the log. This is appended to each time a
trial is run
my $haplog = "haplog.txt";
die "Can't find the log file $haplog\n" unless -e $haplog;
my $lognum = `cat $haplog | wc -l`;
chomp $lognum;
$lognum =~ s/^\s+//; # remove leading whitespace
$lognum =~ s/\s+$//; # remove trailing whitespace

# copy selected output files to unique names at the end
my $hapcopy = "$lognum\_haps.txt";
my $freqcopy = "$lognum\_freqs.txt";
my $poscopy = "$lognum\_pos.txt";
my $wincopy = "$lognum\_win.txt";


# open the legend file to get coordinates of windows as well as windows
details

open FH_LEGEND, $legend_file or die "Could not open legend file
$legend_file.\n";


my @mapfilelines;
my $counter =0;

while (<FH_LEGEND>) {

      chomp;
my ($chr, $rs, $CM, $position) = split /\s+/;

$counter++;

my @linevariables = ($counter, $chr, $rs, $CM, $position);
my $aref = \@linevariables;

push @mapfilelines, $aref;

}
```

```
my $number =0;
my %windows;
for (my $start =0; $start+$winsize <= $mapfilelines[-1][3] ;
$start+=$winshift) {
my @window;
my $wincoor;

for my $g (@mapfilelines) {

if ( ${$g}[3] >= $start and ${$g}[3] < $start+$winsize){
push @window, $g;
}
}
$wincoor = \@window;
$number++;
$windows{$number} = $wincoor;
}


# open outfile to write window coords to
open (FH_POS, ">$posfile") or die "Couldn't open $posfile to write ";



my %windetails;
my $count = 0;

foreach my $key (sort {$a <=> $b} keys %windows) {

my $value = $windows{$key};
my @windowdetails;
my $chrnum;
my $winstart;
my $winend;
my $winref;
my $startrs;
my $startpos;
my $endrs;
my $endpos;
my $markersnum;
my $physicaldis;
my $geneticdis;

my $startcm = ${$value}[0][3];
my $endcm = ${$value}[-1][3];
$chrnum = ${$value}[0][1];
$winstart = ${$value}[0][0];
$winend = ${$value}[-1][0];
$startrs = ${$value}[0][2];
$startpos = ${$value}[0][4];
$endrs = ${$value}[-1][2];
$endpos = ${$value}[-1][4];
$markersnum = $winend - $winstart;
$physicaldis = $endpos - $startpos;
$geneticdis = $endcm - $startcm;

 push @windowdetails, $key;
 push @windowdetails, $chrnum;
 push @windowdetails, $winstart;
 push @windowdetails, $startrs;
 push @windowdetails, $startpos;
```

```
 push @windowdetails, $winend;
 push @windowdetails, $endrs;
 push @windowdetails, $endpos;
 push @windowdetails, $markersnum;
 push @windowdetails, $physicaldis;
 push @windowdetails, $geneticdis;

 $winref = \@windowdetails;
 $count++;
 $windetails{$count} = $winref;
 #$count++;
print FH_POS "window $key is $physicaldis bp in size, has $markersnum
typed markers, with coordinates $startpos -  $endpos, which is between
marker number $winstart $startrs and $winend $endrs.\n";
}
my $first = keys %windetails;
my $second = keys %windows;
print "$first is the number of records in the hash windetails.\n";
print "$second is the number of records in the hash windows.\n";

close FH_LEGEND;

# $| = 1; # don't buffer output

 close FH_POS;


# populate array of haplotypes as strings
my @haps;

# read from haplotype file
local *FH;
open (FH, $infile) or die "Couldn't open $infile to read.\n";
my $full_haplength;
while (<FH>) {
      chomp;
      # skip lines with individuals Ids. ie:skip everything that starts
with WTCCC.
      next if /^TDT/;

      my @temp = split; # splits on whitespace
   my $thishap = join("", @temp); # haplotype as a string like
12112122112121212121
  if (defined $full_haplength) {
    die "Didn't find a haplotype of length $full_haplength" unless
length($thishap) == $full_haplength;
  } else { # infer length to check subsequent lines against from first
line
      $full_haplength = length($thishap);
  }
      push @haps, $thishap;
}
close FH;


my $numchr = scalar @haps;
print "Read $numchr haplotypes of length $full_haplength.\n";

# open main output file
open (FH, ">$hapfile") or die "Couldn't open $hapfile to write ";
my $old_fh = select(FH);
```

```
$| = 1; # don't buffer output to this file
select($old_fh);


# open outfile to write window number, max haplotype frequency, number of
typed markers in the window, window size in bp, window size in cM.
open (FH_WIN, ">$winfile") or die "Couldn't open $winfile to write ";
print FH_WIN
"window\tnumber_of_markers\tmax_hap_freq\tsecond_most_freq_hap\tthird_most
_freq\tforth_most_freq\twindow_size_inbp\twindow_size_inCM\twin_start\twin
_end\twin_coord\n";

# calculation
# loop through the window positions available (pos is starting position, 0
meaning starting from first SNP)
my $max_simhap = 1;
my $max_pos = 0;
my $max_winsize;
my $numpos = 0;
my %numchr_to_freq; # key is number of chromosomes, value is number of
times that number of chromosomes were identical
foreach my $key (sort {$a <=> $b} keys %windetails) {

   #my $poss = $key;
   my $value = $windetails{$key};
   my $chromosome = ${$value}[1];
   my $windowsize = ${$value}[8];
   my $bpdist =  ${$value}[9];
   my $cmdist = ${$value}[10];
   my $pos = ${$value}[2];
   my $startposition = ${$value}[4];
   my $endposition = ${$value}[7];
   my $wincoord =
$chromosome.":".$startposition."_".$chromosome.":".$endposition;

   if ($pos+$windowsize <= $full_haplength) {

   my @hapcounts = count_similar_haps(\@haps, $pos, $windowsize);

   my $markerpos = $key;

   my $maxhaps = 1; # the maximum number of identical haplotypes at this
position
   my $secondfreq;
   my $thirdfreq;
   my $fourthfreq;
   if (scalar @hapcounts) {

      $maxhaps = $hapcounts[0];
      $secondfreq = $hapcounts[1];
      $thirdfreq = $hapcounts[2];
      $fourthfreq = $hapcounts[3];

      if ($hapcounts[0] > $max_simhap) { # keep track of the highest
number of identical haplotypes across positions and the position
         $max_simhap = $hapcounts[0];
         $max_pos = $markerpos;
       $max_winsize = $windowsize;
      }
   }
```

```
   for my $hc (@hapcounts) { # record the frequency of identical
chromosomes (>=2) at this position
      $numchr_to_freq{$hc}++;
   }

   my $outputline = (scalar @hapcounts) ? (join ",", @hapcounts) : '1';

      print FH "$outputline\n";
      $numpos++;


   print FH_WIN
"$numpos\t$windowsize\t$maxhaps\t$secondfreq\t$thirdfreq\t$fourthfreq\t$bp
dist\t$cmdist\t$startposition\t$endposition\t$wincoord\n";
   }
}


close FH;

close FH_WIN;

# print the frequencies to a file for plotting graphs
process_freqs(\%numchr_to_freq,$numchr,$numpos,$freqfile);

# extract and print to file the haplotype strings for the window that had
the highest identical haplotype frequency
extract_best_window(\@haps,$max_pos,$max_winsize,$chkfile);

# print info about this run to the logfile
open FH_LOG, ">>$haplog" or die "Could not open logfile $haplog.\n";
my $time = time-$st;
print FH_LOG "exthap-TDT-cm2.pl log number $lognum\t hapfile
$infile\t$numchr chrom of $full_haplength length\t max num of sim haps
$max_simhap\t at $max_pos\t win size $winsize\t win shift $winshift\t
number of wins $numpos\t finished in $time seconds\n";
close FH_LOG;

print "trial $lognum. Window size is $winsize CM. $numchr haplotypes of
$full_haplength markers ($numpos windows with windowshift $winshift).
$max_simhap of ", scalar @haps, " chromosomes were identical in window
number $max_pos\n";
print "See results:\n";
print "$hapfile - frequencies of haplotypes at each position\n";
print "$freqfile - overall frequencies of haplotypes across all
positions\n";
print "$posfile - window positions and associated recombination rates\n";
print "$winfile - window number, max number of identical haplotypes,
window size in bp and CM\n";
print "Finished in ", $time, " seconds. Log information is in $haplog.\n";
system "cp $hapfile $hapcopy";
system "cp $freqfile $freqcopy";
system "cp $posfile $poscopy";
system "cp $winfile $wincopy";
system "nedit $hapfile $freqfile $posfile $winfile &";


=head1 count_similar_haps
```

```
For a particular window position, examine the substring of haplotypes of a
given size and
count the numbers of identical haplotypes.  We do not include distinct
haplotypes.

=head2 Arguments

 $_[0] : reference to array of haplotypes
 $_[1] : start position (0-based: 0 means beginning of string)
 $_[2] : size of window

Returns empty list if all haplotypes are distinct in this region, or a
list of numbers
like 7, 3, 2, 2 (ordered descending) which represents the number of
identical hits.

=cut

sub count_similar_haps {

   my ($ref_haps, $start, $size) = @_;

   my %subhaps; # hash with key of subhaplotype string, value number of
occurrences

      for my $full_hap (@{$ref_haps}) {
            my $sub_hap = substr($full_hap, $start, $size);
         $subhaps{$sub_hap}++; # will be 1 if not encountered before,
otherwise increment
   }

   # now find those that are not 1
   my @interesting;
   for (values %subhaps) {
      push @interesting, $_ unless $_ == 1;
   }

   return sort {$b<=>$a} @interesting;

} # end sub count_similar_haps

=head1 process_freqs

Calculate the frequency of chromosomes being identical for all number of
chromosomes between 2 and $numchr totalled over all positions.

=head2 arguments

 [0] : a hash of number of chromosomes => freq of being identical
 [1] : total number of chromosomes examined
 [2] : total number of windows used
 [3] : file to write to

=cut

sub process_freqs {

   my ($numchr_to_freq,$numchr,$numpos,$freqfile) = @_;

   open FH, ">$freqfile" or die "Could not open $freqfile.\n";
```

```
   for (my $i = 2; $i <= $numchr; $i++) {
      print FH "$i\t";
   }
   print FH "\n";

   for (my $i = 2; $i <= $numchr; $i++) {

      if (defined $$numchr_to_freq{$i}) {

         # this is the number of times that $i of $numchr chromosomes are
identical
         my $freq = $$numchr_to_freq{$i};
         print FH "$freq\t";

      } else {

         print FH "0\t";
      }

   }
   print FH "\n";
   close FH;

} # end sub process_freqs

=head1 extract_best_window

Extract the haplotype strings for the window that gave the greatest number
of identical haplotypes

=head2 Arguments

 [0] : ref to array of haplotype strings (one element per chromosome)
 [1] : the window position that gives the highest number of identical
haplotypes
 [2] : the size of the window

=cut

sub extract_best_window {

   my ($ref_haps,$max_pos,$windowsize,$chkfile) = @_;

   my @temp;
   for my $full_hap (@{$ref_haps}) {
            my $sub_hap = substr($full_hap, $max_pos, $windowsize);
      push @temp, $sub_hap;
   }

   my @sorted = sort(@temp);

   open FH, ">$chkfile" or die "Could not open $chkfile.\n";
   for my $h (@sorted) {
      print FH "$h\n";
   }
   close FH;

} # end sub extract_best_window
```

Appendix 4: List of Candidate regions of recent adaptive evolution identified by the extended-high-frequency-haplotype analysis as outliers.

Table 1: List of regions identified as outliers in the HapMap YRI sample.

Chromosome	Region start	Region end
1	31556567	33433589
3	86907721	94948713
3	103087318	104285667
4	32485339	34426976
5	28742001	30545765
5	89111186	90513821
7	116185790	118067340
10	57634147	59250452
11	4510238	5995893
11	37097729	39848258
11	83758120	85597492
12	33445564	37798991
12	79150083	80365837
13	45780806	47081488
13	54211160	56561660
13	75210078	76875750
14	44075381	46991264
14	75676344	77064129
15	69757099	72242413
16	35620265	49098715
21	29879196	30682260
21	39093741	39899947
X	60784354	66128220

Table 2: List of regions identified as outliers in the HapMap CEU sample.

Chromosome	Region start	Region end
1	71280907	73209837
2	134815385	139315244
2	152698583	154212372
2	188984759	190690353
3	82170375	85570099
3	88645387	96427710
4	33069411	35043016
5	128843291	132521927
5	143053541	144725999
6	28479325	32787435
7	117007660	119822806
10	68028852	69326346
10	73050857	76184991
11	37801245	39403713
11	89236882	91316765
12	36622847	38720468
12	83925058	85937776
12	86607819	87840431
13	53403182	55351347
14	38232681	39917445
16	66110521	69310080
17	46279132	47949645
17	57722226	59757634
18	23972443	25483998
18	28517165	30324295
18	48147725	49438530
18	61259757	62263986
19	41303124	43375145
20	32368272	34555370
21	28537026	30131445
X	32793093	37015221
X	56100644	66841866

Table 2: List of regions identified as outliers in the HapMap CEU sample.

Chromosome	Region start	Region end
1	71280907	73209837
2	134815385	139315244
2	152698583	154212372
2	188984759	190690353
3	82170375	85570099
3	88645387	96427710
4	33069411	35043016
5	128843291	132521927
5	143053541	144725999
6	28479325	32787435
7	117007660	119822806
10	68028852	69326346
10	73050857	76184991
11	37801245	39403713
11	89236882	91316765
12	36622847	38720468
12	83925058	85937776
12	86607819	87840431
13	53403182	55351347
14	38232681	39917445
16	66110521	69310080
17	46279132	47949645
17	57722226	59757634
18	23972443	25483998
18	28517165	30324295
18	48147725	49438530
18	61259757	62263986
19	41303124	43375145
20	32368272	34555370
21	28537026	30131445
X	32793093	37015221
X	56100644	66841866

Table 3: List of regions exclusive to the Gambian severe malaria cases, with values above two standard deviations (transmitted chromosomes).

Chromosome	Region start	Region end
1	13963744	14635377
1	24450994	25514253
1	151299443	152840623
1	153149865	153931494
1	173888472	175696261
1	220139012	224963515
2	43962338	45017326
2	166377315	168145965
2	234798487	235170750
3	59993366	60324502
3	117934964	118379319
4	70697431	72780842
4	76104443	77488113
4	114593831	116022924
4	137111972	138731721
4	139966470	153459800
4	168022785	169206776
5	98604146	103209350
5	168486834	169404466
5	178112905	178559043
7	6643875	7138246
7	105994486	106900372
7	110076704	111379964
7	129628646	130253907
9	12775073	13500394
9	25732531	26679442
9	77438679	78545080
9	83157154	83938249
10	32430956	33828921
10	36244764	37290394
10	59104312	60334520
12	24949964	25656557
12	41734204	43209601
12	77425651	78635538
13	28413875	29337397
13	62058599	64080366
13	77971249	79148512
13	96857343	97550726
13	107670522	108157835
13	109764350	110091916
14	51311946	51962362
15	63461230	64318418
15	85502521	86262610
16	14378729	15835092
16	69344281	71074515
17	73279820	73775501
18	47185007	48044577
18	48042298	49313161
19	14355473	14806170
20	24265910	28172124

20	58098406	58795868
21	40182104	40514014
22	25625691	25854738

Table 4: List of regions exclusive to the Gambian severe malaria controls, with values above two standard deviations (untransmitted chromosomes).

Chromosome	Region start	Region end
1	18925130	19744951
1	36644986	37217566
1	37206802	38135790
1	39220335	41086073
1	94763951	95860203
1	184148193	185483037
1	209950760	210818616
2	3406189	3804389
2	46547928	47235083
2	121328889	122206162
2	133709244	134559297
2	212203748	212597176
3	67524605	68403369
4	7487097	7688429
4	57318722	61059378
4	62566178	67903349
4	163842473	164833845
5	24880170	25837347
5	29967494	31281091
5	149641346	150439981
6	40880949	41583425
6	69615419	71790469
6	73846091	75593660
6	165855338	166166746
7	13453171	13978945
7	21927694	22873460
7	22697767	24665754
7	25292991	26318339
7	30066932	30711953
7	70649078	72491430
7	142020041	142733566
8	13023771	13535366
9	4364278	4856840
9	16232088	16700870
9	19390351	20076856
9	27994206	29689426
9	31111425	32335581

9	32421306	34413806
10	95239595	95702720
10	113107817	114044591
11	3524620	6871891
11	63976255	67728122
11	68340260	69222519
11	99120957	100375744
12	10839120	11685903
12	111804639	112465947
13	33558808	34360987
13	58702026	60093488
13	111832084	112325637
14	65122654	67056028
14	81465079	82328673
16	10903900	11553435
16	11525159	12069150
16	71941694	72654767
18	44694487	45485063
19	20422200	21828007
19	44329879	45114921
19	59624545	60189667
20	14763778	15317397
20	38882719	40018906
20	59380439	59666382
22	36007876	36301981

Table 5: List of regions identified in the Gambian severe malaria cases and controls, with values above two standard deviations (transmitted and untransmitted chromosomes).

Chromosome	Region start	Region end
1	91912840	94216566
1	112577988	114163349
1	115633140	116436501
1	146487865	147773423
1	171882840	172408553
1	182023083	183738838
1	218283134	219311746
2	31085475	31483437
2	37937323	38537958
2	45474465	45954129
2	52405577	53463251

2	75570435	76557318
2	77992907	79134870
2	106051082	107886535
2	107478040	109104877
2	109163846	111996372
2	122236257	123486075
2	127094744	127487236
2	135592827	137908254
2	157509224	158522037
2	174077451	174749060
2	192896944	195784259
2	205602055	206377484
2	222686780	223177919
2	224004933	225582286
3	3562732	4078926
3	17122967	20347803
3	26362867	27006618
3	26856110	27814871
3	45108910	46327388
3	71332365	71822254
3	71956269	72466185
3	121630554	122754748
3	124342771	125439402
3	130275449	131722422
3	133609225	134639390
4	59478401	61931307
4	68296320	70906581
4	78411154	79640582
4	82841349	83726466
4	100658226	102061562
4	111022234	112363266
4	169926272	170347712
5	53395816	54882855
5	65690748	66586611
5	132637416	133046502
6	26375779	36338113
6	156444174	157490966
7	5701197	6728910
7	18843405	19533430
7	35876184	36325896
7	49397368	51587916
7	111680844	113741168
7	141620986	142566810
8	9011268	9750690
8	25806316	26517618
8	73405227	73998473
9	22234280	23636551
9	35756561	36806714
9	36862314	37721177
9	119529894	120213741
10	21481046	23404726
10	26202772	27707093
10	72002446	72979896
10	76631338	77834518

10	121959101	122502825
11	113625984	116805250
12	12127650	12612565
12	21986082	22836038
12	23016698	23803922
12	32539025	38836591
12	78626764	79648279
12	81034619	82048539
12	86424349	87937645
12	95049481	96309883
13	29345216	29976511
13	29977837	31158884
13	35177556	36228830
15	29143717	30228824
15	51057772	51562962
16	45559127	48183497
16	71475102	71903535
16	76918164	77356110
16	79126074	79681653
17	9948656	10165009
17	15588448	17032630
17	17246812	18503445
17	33914210	34541657
17	60368201	60993743
17	70254715	71109514
18	11068646	11894266
18	13367128	22525080
18	53242291	53928699
18	64343434	64780832
20	23927305	24731483
20	36347834	38090578
20	38303533	38890738
20	39022636	40422758
20	40403504	41081132
20	40912958	41366282
20	45807248	46280573
22	32684271	35857568